ElliCE: Efficient and Provably Robust Algorithmic Recourse via the Rashomon Sets

Bohdan Turbal[1] **Iryna Voitsitska**[2] **Lesia Semenova**[3*]

[1] Princeton University [2] Ukrainian Catholic University [3] Rutgers University
bt4811@princeton.edu, voitsitska.pn@ucu.edu.ua, lesia.semenova@rutgers.edu

Abstract

Machine learning models now influence decisions that directly affect people's lives, making it important to understand not only their predictions, but also how individuals could act to obtain better results. Algorithmic recourse provides actionable input modifications to achieve more favorable outcomes, typically relying on counterfactual explanations to suggest such changes. However, when the Rashomon set – the set of near-optimal models – is large, standard counterfactual explanations can become unreliable, as a recourse action valid for one model may fail under another. We introduce ElliCE, a novel framework for robust algorithmic recourse that optimizes counterfactuals over an ellipsoidal approximation of the Rashomon set. The resulting explanations are provably valid over this ellipsoid, with theoretical guarantees on uniqueness, stability, and alignment with key feature directions. Empirically, ElliCE generates counterfactuals that are not only more robust but also more flexible, adapting to user-specified features constraints while being substantially faster than existing baselines. This provides a principled and practical solution for reliable recourse under model uncertainty, ensuring stable recommendations for users even as models evolve.

1 Introduction

When an algorithmic decision denies someone a loan, a job, or insurance coverage, a natural question follows: *What could I change to obtain a better outcome next time?* Algorithmic recourse answers this question by providing concrete, actionable changes that could lead to a more favorable decision. A common way to generate such recommendations is through counterfactual explanations – small modifications to an individual's features that flip the model's prediction. Yet, even when the recommendation looks specific (e.g. "increase your income by $5000"), one must ask: *Would that same change still work tomorrow if the institution retrains or replaces its model?* or *How stable are these suggestions across equally good models that explain the data in different ways?*

Most existing counterfactual generation methods [25, 46, 50, 52, 54, 57, 61, 66] implicitly assume that the underlying model is fixed and perfectly known. In practice, models evolve: banks regularly retrain risk predictors, healthcare institutions update diagnostic classifiers, and regulators may require model re-validation under new privacy or transparency constraints. Small shifts in data or regularization can result in very different-yet-equally-accurate models. This phenomenon, known as the Rashomon effect [8, 18, 23, 56, 58], implies that many distinct predictors achieve nearly optimal performance. In such settings, a counterfactual valid for one model can fail under another, undermining the reliability and consistency of algorithmic recourse.

*The majority of this work was conducted while Bohdan was at Taras Shevchenko National University of Kyiv and Lesia at Microsoft Research NYC.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Recent approaches have attempted to produce robust counterfactuals, meaning counterfactuals that are valid under small parameter perturbations or across predefined ensembles [17, 22, 27, 34, 36, 37, 41, 63]. However, these methods either rely on heavy-weight mixed-integer solvers, restrict robustness to local neighborhoods around a single model, or lack formal guarantees of validity across the full space of near-optimal solutions known as the Rashomon set. None of them directly leverages the geometry of this Rashomon set itself.

We introduce ElliCE, an efficient and provably robust framework for algorithmic recourse that works over an ellipsoidal approximation of the Rashomon set. By modeling the space of near-optimal models as an ellipsoid derived from the curvature (Hessian) of the loss landscape, ElliCE reformulates robust counterfactual generation as a tractable convex optimization problem. The resulting counterfactuals are valid for every model inside the ellipsoid, ensuring that a user's recommended action remains meaningful even if the deployed model is replaced by another equally accurate one from the approximated Rashomon set.

Our contributions are fourfold: (1) *Theoretical foundation*. We derive a closed-form expression for the worst-case prediction, which allows us to formulate the robust recourse problem as a convex optimization and establish formal guarantees of validity, uniqueness, and stability for ElliCE's counterfactuals. (2) *Geometric intuition*. We show that ElliCE's robustness term connects the counterfactual's stability with the importance of the features it modifies as the optimization naturally aligns recourse directions with the principal curvature axes of the loss landscape. (3) *Actionability*. ElliCE supports feature-level constraints, such as sparsity constraints, immutable or range-restricted attributes, allowing users to generate realistic, actionable recourse tailored to specific application or user settings. (4) *Empirical validation*. Across multiple high-stakes tabular datasets, ElliCE achieves higher robustness and remains one to three orders of magnitude faster than competing baselines, while maintaining proximity and plausibility.

Ultimately, ElliCE looks at algorithmic recourse through the lens of model multiplicity. Instead of relying on a single model's decision boundary, it offers explanations that stay consistent across many models that fit the data almost equally well. This perspective treats the Rashomon Effect not as a flaw to eliminate, but as an inherent source of uncertainty to account for, leading to stable recourse in the presence of model diversity.

2 Related works

Rashomon Effect. The Rashomon Effect, a term popularized by Breiman [8] in the context of machine learning, describes the phenomenon where multiple distinct models can achieve near-optimal empirical risk (these models form a Rashomon set). This effect is also referred to as model multiplicity [5, 48]. The existence of the Rashomon set has implications for the trustworthiness and reliability of machine learning systems, influencing feature importance [14, 15, 20, 51], fairness [13, 45, 49], the existence of simple yet accurate models [6, 58, 59] to name a few. Significant research has focused on measuring and characterizing the Rashomon set for different model classes [29, 31, 32, 68, 70]. Our work leverages insights into the geometry of the Rashomon set, explored by works like Donnelly et al. [16], Zhong et al. [70], but applies them to the distinct challenge of generating robust algorithmic recourse across this set.

Counterfactual Explanations. Counterfactual Explanations (CEs) have emerged as a prominent tool for providing algorithmic recourse. Numerous approaches exist for generating CEs. Proximity-based methods aim for counterfactuals requiring minimal feature space perturbations [9, 50, 65, 66]. Sparsity techniques prioritize modifying the fewest features possible to enhance actionability [25, 61], while some methods attempt to balance both objectives [46]. Another research direction emphasizes plausibility, ensuring generated CEs represent realistic instances by constraining them to the data manifold, for example, using guidance from generative models [39, 52, 53], encoding feasibility rules [40], or tracing density-aware paths [54]. Recent extensions also incorporate temporal reasoning [11] and fairness objectives [4, 43, 69]. A key limitation across these approaches (which ElliCE directly addresses) is the assumption of a fixed, perfectly known predictive model, as counterfactuals constructed near a specific decision boundary can become unstable under model updates or perturbations.

Robustness to Local Model Perturbations. Building upon the limitation of fixed models, one line of work has focused specifically on achieving robustness against small, local changes or per-

turbations in the model's parameters. For instance, ROAR [63] optimizes CEs considering local Δ-perturbations of the model. Jiang et al. [33] introduced Δ-robustness, a formal metric to assess CE validity under bounded parameter perturbations in neural networks, with subsequent works developing provably robust MILP-based methods [34, 36]. While these methods offer formal guarantees for Δ-robustness, MILP-based approaches can face scalability challenges, and the focus is generally on local parameter stability rather than the broader implications of the Rashomon effect.

Robustness under the Rashomon Effect. A growing body of work addresses counterfactual robustness under model multiplicity, aligning closely with the Rashomon Effect. Several approaches evaluate stability across predefined sets or ensembles of models, introducing heuristic stability measures (e.g., T:Rex [27] and RobX [17]), probabilistic frameworks [22, 41], or guarantees under specific norms and conditions like distribution shift [24, 42, 44]. Foundational work by Pawelczyk et al. [53] conceptually linked the Rashomon Effect to counterfactuals, though primarily enhancing input perturbation robustness. More recent methods use argumentative ensembling [37] or aggregate explanations across AutoML-generated sets [10] to handle model multiplicity.

Our work takes a distinct approach. Rather than relying on ensemble agreement, heuristic stability metrics, local perturbations, or argumentative aggregation, ElliCE leverages the local geometry of the Rashomon set, approximated by an ellipsoid, to derive theoretically grounded, robust recourse valid across all models within the approximation.

3 Background and Notation

Dataset and hypothesis space. Consider n i.i.d. samples $\mathcal{S}_n = \{\mathbf{z}_i = (\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and $y_i \in \mathcal{Y} = \{0, 1\}$ are generated from an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{Y}_{pred}$ be an output space, where $\mathcal{Y}_{pred} \subseteq \mathbb{R}$ for scores (logits) or $\mathcal{Y}_{pred} \subseteq [0, 1]$ for probabilities. Then $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is a hypothesis space of functions $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}_{pred}$, parameterized by a vector $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. For example, $\mathcal{F}$ can represent linear models or multilayer perceptrons. We denote a specific function by $f_{\boldsymbol{\theta}}$. As our analysis focuses on the parameter space Θ, we will often refer to the model directly by its parameter vector $\boldsymbol{\theta}$.

Loss and objective function. Let $\phi : \mathcal{Y}_{pred} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function. In this work, we consider logistic loss $\phi(f_{\boldsymbol{\theta}}(\mathbf{x}), y) = -[y \log(\sigma(f_{\boldsymbol{\theta}}(\mathbf{x}))) + (1 - y) \log(1 - \sigma(f_{\boldsymbol{\theta}}(\mathbf{x})))]$, which is applied to the model's raw score (logit), $s = f_{\boldsymbol{\theta}}(\mathbf{x})$, where $\sigma(s) = \frac{1}{1+\exp(-s)}$ is the sigmoid function. However, our results generalize to other convex losses. The true risk is the expected loss $J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}}[\phi(f_{\boldsymbol{\theta}}(\mathbf{x}), y)]$ that we approximate with the empirical risk, which is the average loss, $\hat{J}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \phi(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$. We also define an ℓ_2-regularized objective function: $\hat{L}(\boldsymbol{\theta}) = \hat{J}(\boldsymbol{\theta}) + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$, where $\lambda \geq 0$ is the regularization strength. The empirical risk minimizer (ERM) is $\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \hat{L}(\boldsymbol{\theta})$. When $\lambda = 0$, the ERM is $\hat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \hat{J}(\boldsymbol{\theta})$.

Rashomon set. Following [21, 58, 67], we define the ϵ-Rashomon set within the parameter space Θ as the set of parameter vectors whose corresponding models $f_{\boldsymbol{\theta}}$ have objective value close to the minimum:

$$\mathcal{R}(\epsilon) \coloneqq \{\boldsymbol{\theta} \in \Theta : \hat{L}(\boldsymbol{\theta}) \leq \hat{L}(\hat{\boldsymbol{\theta}}) + \epsilon\},$$

where $\epsilon \geq 0$ is the Rashomon parameter defining the allowable tolerance in objective compared to the ERM, $\hat{L}(\hat{\boldsymbol{\theta}})$. It is typically a small value. The existence of the Rashomon set with multiple, distinct parameter vectors $\boldsymbol{\theta}$ (corresponding to potentially distinct functions $f_{\boldsymbol{\theta}}$) achieving similar performance implies that different underlying logic (how features contribute to predictions) can explain the data equally well. It is important to be aware of this variability among near-optimal models when generating explanations for individual predictions, as different models in $\mathcal{R}(\epsilon)$ might suggest different ways an outcome could be changed.

Counterfactual explanations. Let $g : \mathcal{Y}_{pred} \to \{0, 1\}$ be the decision function that converts a model's score output $s = f_{\boldsymbol{\theta}}(\mathbf{x})$ to a final binary class label by applying a threshold t, such that $g(s) = \mathbf{1}[s \geq t]$. For an ERM $\hat{\boldsymbol{\theta}}$ and for an input vector $\mathbf{x}_0$ with prediction $g(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_0)) = \hat{y}_0$, a counterfactual explanation $\mathbf{x}_c$ is a data point such that its predicted class is the opposite, i.e., $g(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_c)) = 1 - \hat{y}_0$. The set of all counterfactual explanations for $\mathbf{x}_0$ under the model $\hat{\boldsymbol{\theta}}$ and decision function g is defined as:

$$\mathcal{C}(\mathbf{x}_0, \hat{\boldsymbol{\theta}}) = \left\{\mathbf{x}_c \in \mathcal{X} : g(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_c)) = 1 - g(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_0))\right\}.$$

For instance, in a credit loan application scenario, if an applicant $\mathbf{x}_0$ is denied a loan (e.g., $g(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_0)) = 0$), a counterfactual explanation $\mathbf{x}_c$ would be a modified version of their application details (e.g., increased income, reduced debt) such that the model predicts approval, $g(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_c)) = 1$. While many such $\mathbf{x}_c$ might exist, practical algorithmic recourse aims to find explanations that require minimal change for the user. This means finding the "closest" counterfactual: $\mathbf{x}_c^* = \arg\min_{\mathbf{x}_c \in \mathcal{C}(\mathbf{x}_0, \hat{\boldsymbol{\theta}})} \nu(\mathbf{x}_c, \mathbf{x}_0)$, where $\nu(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a defined distance function or cost metric.

Distance Metrics. In our framework, we primarily focus on the two distance metrics for generating actionable and interpretable counterfactuals: ℓ_2 or Euclidean distance and mixed distance ℓ_{mix}. Note that ℓ_2 is a natural geometric measure of proximity, that penalizes large differences in any feature, $\ell_2(\mathbf{x}_c, \mathbf{x}_0) = \|\mathbf{x}_c - \mathbf{x}_0\|_2^2 = \sum_{j=1}^{d}(x_{cj} - x_{0j})^2$. For practical applications where features have different natures (continuous and categorical), one can also consider the mixed distance ℓ_{mix}, inspired by Gower's distance. Assuming that the data are standardized, it is defined as: $\ell_{mix}(\mathbf{x}_c, \mathbf{x}_0) = \sqrt{\sum_{j \in \mathcal{I}_{cont}}(x_{cj} - x_{0j})^2 + \sum_{j \in \mathcal{I}_{cat}} w_j \mathbf{1}[x_{cj} \neq x_{0j}]}$, where $\mathcal{I}_{cont}$ and $\mathcal{I}_{cat}$ denote the sets of continuous and categorical feature indices respectively, $\mathbf{1}[\cdot]$ is the indicator function, and w_j are optional weights reflecting the cost of changing feature j. We use ℓ_2 distance for our theoretical analysis in the subsequent sections.

Next, we describe our approximating framework and outline the optimization process.

4 A Framework for Robust Recourse over the Rashomon Set

We focus our theoretical analysis on linear predictors of the form $f_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x}$. However, the same methodology applies in the final embedding space of multilayer perceptrons (MLPs) by writing the model as $f_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top h(\mathbf{x})$, where $h(\mathbf{x})$ is the penultimate-layer embedding and $\boldsymbol{\theta}$ are the last-layer parameters. We freeze $h(\cdot)$ and apply the same ellipsoidal procedure to $\boldsymbol{\theta}$ as in the linear case (equivalently, replace $\mathbf{x}$ by $h(\mathbf{x})$ in the formulas below).

Approximated Rashomon set. For certain objectives, such as ℓ_2-regularized mean-squared error on linear models, the Rashomon set is exactly an ellipsoid in the parameter space [58]: $\mathcal{R}(\epsilon) = \{\boldsymbol{\theta} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top (X^\top X + \lambda I_p)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq \epsilon\}$, where $X \in \mathbb{R}^{n \times d}$ is the data matrix, whose i-th row is the feature vector $\mathbf{x}_i^\top$, I_p is an identity matrix of size $p \times p$, and $\lambda \in \mathbb{R}_+$ is the regularization parameter. Because mean-squared error provides a local quadratic approximation to other convex losses, the exact ellipsoidal form of its Rashomon set serves as strong motivation for the Rashomon set approximation. Building on this and on similar geometric intuition [70], we approximate the ϵ-Rashomon set with an ellipsoid defined by the local geometry of the loss landscape:

$$\hat{\mathcal{R}}(\epsilon) = \{\boldsymbol{\theta} : \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top H (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq \epsilon\},$$

where $H = X^\top W X + \lambda I_p$ is the Hessian of the ℓ_2-regularized loss function, evaluated at $\hat{\boldsymbol{\theta}}$. For logistic loss, W is an $n \times n$ diagonal matrix of weights where $w_{ii} = \sigma(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_i))(1 - \sigma(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_i)))$. Recall from Section 3 that $\sigma(\cdot)$ is the sigmoid function.

The Hessian matrix H of the regularized objective $\hat{L}(\boldsymbol{\theta})$ is strictly positive definite. This is because it is the sum of the positive semidefinite (PSD) Hessian from the convex logistic loss and the positive definite (PD) Hessian from the ℓ_2 regularization term (λI_p), assuming $\lambda > 0$. A positive definite Hessian is important for our framework, as it guarantees the approximating ellipsoid is well-defined and bounded, and ensures that H is invertible for our closed-form solution.

In cases where the unregularized risk $\hat{J}(\boldsymbol{\theta})$ is minimized (e.g., for neural networks), the resulting Hessian is only guaranteed to be PSD and may be singular. For these models, we ensure positive definiteness in practice by adding a small stabilization term, αI_p, $\alpha > 0$, to the computed Hessian, which is a standard technique to guarantee invertibility.

Optimization. To find a robust counterfactual explanation, we want to compute an explanation $\mathbf{x}_c$ that is closest to the original point $\mathbf{x}_0$ while ensuring that its predicted outcome is above a target threshold t for all models within the approximated Rashomon set. In other words, for a given $\mathbf{x}_0$, we look for a minimally modified (measured in some distance; we will use ℓ_2 here) $\mathbf{x}_c$, such that its predicted outcome achieves t even when evaluated by the least favorable model $\boldsymbol{\theta}$ within the

approximated Rashomon set $\hat{\mathcal{R}}(\epsilon)$. Formally, this requirement leads to the following optimization problem:

$$\min_{\mathbf{x}_c} \quad \|\mathbf{x}_c - \mathbf{x}_0\|_2^2 \qquad \text{s.t.} \quad \min_{\boldsymbol{\theta} \in \hat{\mathcal{R}}(\epsilon)} \boldsymbol{\theta}^\top \mathbf{x}_c \geq t. \tag{1}$$

The inner minimization problem admits a closed-form solution, as we show next in Theorem 1. By reformulating the problem in this way, we get a tractable optimization framework that supports more efficient computation and analytical analysis of solution properties.

Theorem 1 (Closed-form solution)**.** *For positive-definite Hessian H, the inner minimization problem over the ellipsoid-approximated Rashomon set $\hat{\mathcal{R}}(\epsilon)$ has the closed-form solution $\min_{\boldsymbol{\theta} \in \hat{\mathcal{R}}(\epsilon)} \boldsymbol{\theta}^\top \mathbf{x}_c = \hat{\boldsymbol{\theta}}^\top \mathbf{x}_c - \sqrt{2\epsilon\, \mathbf{x}_c^\top H^{-1} \mathbf{x}_c}$. Moreover, for a given $\mathbf{x}_c$, the worst-case model $\boldsymbol{\theta}_{worst}(\mathbf{x}_c)$ that achieves this minimum is: $\boldsymbol{\theta}_{worst}(\mathbf{x}_c) = \hat{\boldsymbol{\theta}} - \sqrt{2\epsilon}\, \frac{H^{-1}\mathbf{x}_c}{\sqrt{\mathbf{x}_c^\top H^{-1} \mathbf{x}_c}}$.*

We prove Theorem 1 in Appendix A.1. As a direct consequence of Theorem 1, we obtain a practical criterion for verifying the robustness of a potential counterfactual. Specifically, since the theorem provides an explicit characterization of the output generated by the least favorable model $\boldsymbol{\theta} \in \hat{\mathcal{R}}(\epsilon)$ for a given $\mathbf{x}_c$, we can immediately determine if this $\mathbf{x}_c$ achieves the target t across the entire set as we show in the following corollary.

Corollary 1. *A given counterfactual explanation $\mathbf{x}_c$ is robust with respect to all models in the ellipsoid-approximated Rashomon set $\hat{\mathcal{R}}(\epsilon)$ against a target score t if and only if: $\hat{\boldsymbol{\theta}}^\top \mathbf{x}_c - \sqrt{2\epsilon\, \mathbf{x}_c^\top H^{-1} \mathbf{x}_c} \geq t$.*

By substituting the closed-form solution from Theorem 1 into the original optimization problem (1), the robust counterfactual optimization problem becomes:

$$\min_{\mathbf{x}_c} \quad \|\mathbf{x}_c - \mathbf{x}_0\|_2^2 \qquad \text{s.t.} \quad \hat{\boldsymbol{\theta}}^\top \mathbf{x}_c - \sqrt{2\epsilon\, \mathbf{x}_c^\top H^{-1} \mathbf{x}_c} \geq t. \tag{2}$$

The resulting problem is a quadratically constrained quadratic program (QCQP), which is a class of tractable convex optimization problems. We solve it efficiently using a gradient-based method. Leveraging the formulation (2), we implement two approaches for generating counterfactuals: a search-based method for generating data-supported counterfactuals lying on the data manifold, and a continuous optimization method for exploring potentially novel non-data supported solutions.

Continuous CE generation. For non-data supported counterfactuals, we solve the convex optimization problem in Equation (2) using a gradient-based approach for both linear models and multilayer perceptrons. This method directly optimizes for a counterfactual $\mathbf{x}_c$ in the input space. For neural networks, the process is guided by the worst-case model $\boldsymbol{\theta}_{worst}(\mathbf{x}_c)$ identified in the final layer's embedding space using Theorem 1, with the resulting gradients mapped back to the input features. The full details of this procedure are available in Appendix B.4.

Data-supported CE generation. For practical applications where counterfactuals should remain on the data manifold, we generate data-supported explanations based on the training set. Specifically, we evaluate the robust logit $\hat{\boldsymbol{\theta}}^\top \mathbf{x}_i - \sqrt{2\epsilon \mathbf{x}_i^\top H^{-1} \mathbf{x}_i}$ for each training data point $\mathbf{x}_i$ using Theorem 1. Then, we compute the subset S_{stable} by filtering out points where this robust prediction exceeds the target threshold t. Finally, we use k-d tree nearest neighbor search within S_{stable} to identify the points closest to the input point $\mathbf{x}_0$ in terms of defined distance (for example, ℓ_2), which gives us a counterfactual that is both robust and lies on the data manifold.

The continuous approach offers flexibility by exploring the entire feature space for new solutions, while the data-supported approach guarantees plausibility by restricting solutions to observed examples. We evaluate the performance of both approaches in Section 6 and focus on theoretical guarantees of our framework next.

5 Theoretical Guarantees of ElliCE Counterfactuals

In this section, we explore key theoretical properties of the counterfactual explanations generated under our framework. Note that we use ℓ_2 distance as target distance between $\mathbf{x}_0$ and $\mathbf{x}_c$. We show that the counterfactual explanations generated by our method are valid, unique, stable, and align

with important directions in the feature space. We focus on each of these properties separately and proofs of theorems provided in this section are in Appendix A.2.

Validity. By explicitly optimizing for the worst-case model $\boldsymbol{\theta}_{worst}$ within the defined ellipsoid, any counterfactual $\mathbf{x}_c$ generated by ElliCE is, by construction, valid for all models in the approximated Rashomon set. This inherent validity ensures that the provided recourse is faithful, regardless of which model from the approximated Rashomon set was selected.

Uniqueness. By Theorem 2, that we state next, any solution $\mathbf{x}_c$ to the modified optimization problem (2) is unique. Because our objective is strictly convex and the approximated Rashomon set is characterized as an ellipsoid, for a given $\mathbf{x}_0$, there can never be two distinct counterfactuals at the same ℓ_2 distance from the original $\mathbf{x}_0$. In practical terms, this uniqueness guarantees that ElliCE provides a single solution for a given input and desired robustness level. This directly addresses and resolves "explanation multiplicity" [26], where multiple, distinct explanation paths might exist for a single input (at least for ℓ_2 distance).

Theorem 2 (Uniqueness)**.** *If a solution $\mathbf{x}_c$ to the optimization problem* (2) *exists, then $\mathbf{x}_c$ is unique.*

Stability. The input data $\mathbf{x}_0$ is often subject to noise or minor variations. A desirable property is that such small changes in the input do not lead to drastically different counterfactuals. Our framework ensures this stability. Theorem 3 formally states that the process of generating robust counterfactuals is Lipschitz continuous with a constant of 1. This means that if the original input $\mathbf{x}_0$ is perturbed by a small amount $\boldsymbol{\delta}$ to become $\mathbf{x}'_0$, the resulting robust counterfactual $\mathbf{x}'_c$ will not deviate from the original counterfactual $\mathbf{x}_c$ by more than the magnitude of the initial perturbation $\|\boldsymbol{\delta}\|_2$. This property guarantees the reliability of the explanations.

Theorem 3 (Stability)**.** *Given an input $\mathbf{x}_0$, let $\mathbf{x}_c$ be the robust counterfactual solution for $\mathbf{x}_0$. If the input is perturbed to $\mathbf{x}'_0 = \mathbf{x}_0 + \boldsymbol{\delta}$, where $\boldsymbol{\delta} \in \mathbb{R}^d$, and $\mathbf{x}'_c$ is the robust counterfactual solution for $\mathbf{x}'_0$, then $\|\mathbf{x}_c - \mathbf{x}'_c\|_2 \leq \|\boldsymbol{\delta}\|_2$.*

Alignment with Important Feature Directions. An insightful explanation should not only provide a path to a different outcome but also highlight which features are most critical in achieving that change, particularly under model uncertainty. The robustness penalty term, $C_{rob}(\epsilon, \mathbf{x}_c) = \sqrt{2\epsilon \mathbf{x}_c^\top H^{-1} \mathbf{x}_c}$, plays a key role in this alignment. Theorem 4 formalizes the intuition that to minimize this penalty (and thus find an efficient robust counterfactual), the recourse direction $\mathbf{x}_c$ should align with directions in the feature space that are most sensitive or influential, as captured by the eigenvectors of the Hessian matrix H. Specifically, under certain conditions, the penalty is minimized when the counterfactual aligns with the leading eigenvector of H, which often corresponds to the direction of greatest sensitivity. This encourages the counterfactual to suggest changes along features that have a significant impact, making the explanation more informative.

Theorem 4 (Alignment with Important Feature Directions)**.** *Define the robustness penalty as $C_{rob}(\epsilon, \mathbf{x}_c) = \sqrt{2\epsilon \mathbf{x}_c^\top H^{-1} \mathbf{x}_c}$ for a symmetric positive definite Hessian H. Let λ_1 be the largest eigenvalue of H with corresponding eigenvector $\mathbf{q}_1$, and assume that λ_1 is unique. Then the robustness penalty term $C_{rob}(\epsilon, \mathbf{x}_c)$ is minimized when the counterfactual vector $\mathbf{x}_c$ is aligned with the eigenvector $\mathbf{q}_1$.*

Price of robustness. Previous literature has observed the trade-off between robustness and proximity [22]. Indeed, intuitively, increasing robustness and ensuring validity across a larger set of potential models may require more changes to the input features, effectively increasing the proximity. This implies a "cost" for greater robustness that Theorem 5 formalizes.

Theorem 5 (Robustness-Proximity Trade-off)**.** *For an input $\mathbf{x}_0$ such that $\hat{\boldsymbol{\theta}}^\top \mathbf{x}_0 \leq t$, where $\hat{\boldsymbol{\theta}}$ is ERM, let $\mathbf{x}_c^*(\epsilon)$ be the optimal robust counterfactual for a given robustness level $\epsilon > 0$, and let $\nu(\epsilon) = \|\mathbf{x}_c^*(\epsilon) - \mathbf{x}_0\|_2^2$ be its ℓ_2 distance from $\mathbf{x}_0$. If $\nu(\epsilon_1) > 0$ and $\mathbf{x}_c^*(\epsilon_1) \neq \mathbf{0}$, then for any two robustness levels $0 < \epsilon_1 < \epsilon_2$, $\nu(\epsilon_1) < \nu(\epsilon_2)$.*

The practical impact of this trade-off is significant. Overly robust counterfactuals may become distant and unactionable, while insufficient robustness compromises recourse reliability under model shifts. This underscores the need for methods that efficiently explore this trade-off by achieving substantial robustness with reasonable proximity – a goal that ElliCE effectively meets.

When applying our theoretical results to MLPs, *the validity guarantee is fully preserved in the input space*, which is a key result. The formal guarantees for uniqueness (Theorem 2), stability

(Theorem 3), and the robustness-proximity trade-off (Theorem 5), however, depend on the convexity of the feasible set (see proof of Theorem 2). While this convexity is guaranteed in the embedding space $h(\mathbf{x})$, the nonlinear mapping from the input space ($\mathbf{x} \mapsto h(\mathbf{x})$) means it is not guaranteed to hold there. This distinction highlights a fundamental challenge for robust recourse in deep models and underscores that extending these formal guarantees to the input space is a promising direction for future work. Nonetheless, these theorems provide a principled geometric foundation for our approach and hold for linear models and embedding spaces. Next, we present empirical results showing that ElliCE's performance is consistent with its theoretical guarantees.

6 Evaluation Pipeline and Experimental Results

In our evaluation pipeline, we work with the hypothesis space of linear models and multi-layer perceptrons. However, our results can be extended to other hypothesis spaces that can be optimized with gradient descent, such as neural additive models [1]. In this section, we empirically show that ElliCE is faster and more robust as compared to other methods that produce robust counterfactuals Please see Appendix B for additional details and results.

Datasets. We consider nine datasets from high-stakes decision domains such as lending (Australian Credit [55], FICO [19], German Credit [28], Banknote [47]), healthcare (Parkinson's [62], Diabetes [60]), and recidivism (COMPAS [2]), as well as benchmark datasets (Wine Quality [12], Extended Iris [3]). Please see Table 3 for detailed dataset descriptions and preprocessing notes. We used datasets with predominantly categorical features (FICO, Australian Credit, COMPAS, German Credit, Diabetes) for data-supported CE generation, and datasets with continuous features (Diabetes, Parkinson's, Banknote, Iris, and Wine Quality) for continuous methods. We balanced the datasets, standardized continuous features, and, for some datasets, dropped rows with missing values.

Baselines. We compare ElliCE to other methods that are designed to generate robust counterfactual explanations, such as T:Rex, Interval Abstractions (we refer to it as Delta-robustness [33]), PROPLACE, and ROAR. *T:Rex* [27] generates robust counterfactuals for neural networks using a Stability measure that depends on variance. It quantifies robustness to naturally occurring model changes, providing probabilistic validity guarantees. It is a successor of RobX [17], which targets tree-based ensembles. *Interval Abstractions* [36] ensures that counterfactuals are robust to bounded changes in model parameters (weights and biases). It uses interval neural networks and mixed-integer linear programming. *PROPLACE* [35] formulates counterfactual generation as a bi-level robust optimization problem: it enforces plausibility by restricting solutions to the convex hull of realistic samples and uses interval bounds on neural networks to ensure robustness. *ROAR* [64] optimizes counterfactual validity under bounded model parameter perturbations using a robustness-constrained loss formulation. Most implementations of our baselines follow Jiang et al. [38].

Evaluators. Precisely computing the entire Rashomon set for the hypothesis spaces that we consider is intractable. Therefore, to evaluate the robustness and validity of counterfactual explanations generated by ElliCE and the baselines, we rely on established techniques that approximate or characterize this set. These approaches generate diverse collections of near-optimal models, each serving as a proxy for the true Rashomon set. Our evaluators include: *Random Retrain*, which retrains models multiple times with different random seeds to capture procedural variability. *Rashomon Dropout* [32], which applies binary dropout masks to a single trained neural network's weights during inference, creating an ensemble of thinned sub-models. *Adversarial Weight Perturbation (AWP)* [30], which generates diverse models from an initially trained model by applying small perturbations to its weights. We define the objective tolerance (Rashomon parameter) for the evaluators as $\varepsilon_{\text{target}}$, which is distinct from ϵ. This separation ensures that the Rashomon set used for evaluation is controlled independently from the robustness tolerance ϵ used by ElliCE.

Metrics. We evaluate the generated counterfactual explanations based on four metrics: validity, proximity, robustness, and plausibility. *Validity* measures whether a generated counterfactual $\mathbf{x}_c$ for a given input $\mathbf{x}_0$ successfully achieves the desired outcome c when evaluated on the original model f_{baseline} for which it was generated, Validity $= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[f_{\text{baseline}}(\mathbf{x}_{ci}) = c]$. *Proximity* measures the closeness of a counterfactual $\mathbf{x}_c$ to the original instance $\mathbf{x}_0$. We primarily report the ℓ_2 distance: $\|\mathbf{x}_c - \mathbf{x}_0\|_2$. Lower values indicate less change required and are thus better. *Plausibility* checks whether the generated counterfactuals lie in realistic regions of the feature space. Our data-supported counterfactuals are inherently plausible, as they lie on the data manifold. For con-

tinuous approach, because ElliCE enforces robustness by pushing counterfactuals away from the decision boundary, the resulting counterfactuals tend to shift toward higher-density regions of the target class. Nevertheless, we evaluate plausibility using the Local Outlier Factor (LOF) [35], a standard outlier-detection metric. LOF values close to 1 indicate high plausibility, whereas larger values suggest the counterfactual is in a low-density region. *Robustness* computes whether the generated counterfactual $\mathbf{x}_c$ remains valid (i.e., still achieves the desired outcome c) for all models within an evaluator ensemble $\tilde{\mathcal{R}}(\varepsilon_{\text{target}})$. Total is calculated as the average across all n counterfactual points: $\text{Robustness} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left[\forall f_{\boldsymbol{\theta}} \in \tilde{\mathcal{R}}(\varepsilon_{\text{target}}), f_{\boldsymbol{\theta}}(\mathbf{x}_{c_i}) = c\right]$. A higher robustness score (closer to 1) is better, indicating that more counterfactual explanations are robust to model changes.

Experimental Setup. For evaluators, we define a target multiplicity tolerance globally in range $\varepsilon_{\text{target}} \in [0, 0.1]$. We provide discussion on how to choose ElliCE's ϵ in Appendix B. For every dataset, we performed 4-fold stratified cross-validation. Within each fold, the training data are further split into 80% for training and 20% for validation. The procedure within each inner fold is as follows: (1) We train a base model f_{baseline}, which serves as a reference model for all counterfactual generation methods. (2) Using f_{baseline} as a reference (if required by the evaluation method), we generate $\varepsilon_{\text{target}}$-Rashomon set. (3) Multiplicity parameters (ϵ for ElliCE, δ for Delta Robustness, ROAR and PROPLACE, or τ for T:Rex) for each baseline are tuned via grid search on the validation set with a goal of maximizing validity. We allocate approximately the same amount of time for each method to tune its parameters with a hard maximum of 8 hours per method per data fold (as a result, we could not run PROPLACE for Parkinsons dataset). (4) Final performance metrics are reported on the held-out split of the outer fold. Note that due to our tuning procedure, we expect high validity metric for ElliCE and baselines. Indeed, for data-supported methods validity is consistently 100% across datasets, so we do not report it.

We conducted experiments on logistic regression and multilayer perceptrons. Consistent with prior work [34, 63], we focus on generating counterfactuals that change predicted labels from 0 to 1. Linear models are trained using Scikit-learn's LBFGS solver with an ℓ_2 penalty (regularization parameter 0.001). MLPs are trained with the Adam optimizer (learning rate 0.001), early stopping, and ℓ_2 regularization parameter 0.001. For evaluation, we generate one counterfactual per method for each data point in the held-out set. Each counterfactual is then evaluated against the three evaluators (Random Retrain, Rashomon Dropout, AWP). The exact construction algorithms for these evaluators are described in Appendix B.2. Reported metrics are averaged across data points and folds, with plots displaying the mean and standard error.

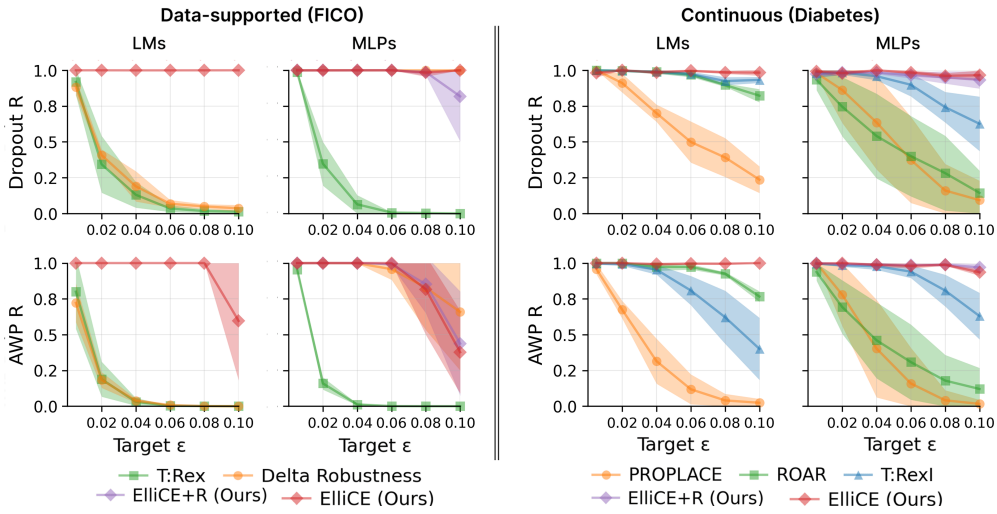

Figure 1: Robustness evaluation of ElliCE against baselines. The plot displays the robustness metric (y-axis) as a function of the target robustness level $\varepsilon_{\text{target}}$ for the evaluators (x-axis). ElliCE consistently outperforms all baselines across all robustness levels. See Appendix B for more figures. With ElliCE+R for MLPs we apply additional regularization to the Hessian, $\lambda = 0.1$, instead of 0.001.

Table 1: Performance of counterfactual methods on MLPs. For evaluators, we set $\varepsilon_{\text{target}}$ to 10% of the training objective ($\varepsilon_{\text{target}} = 0.1 \times \hat{L}(f_{\text{baseline}})$). **R** here stands for Robustness, **L2** for proximity, and PROP stands for PROPLACE. See Appendix B for results on other datasets.

Data	Method	Evaluation Metric					
		Retrain		Dropout Rashomon		AWP	
		R↑	**L2↓**	**R↑**	**L2↓**	**R↑**	**L2↓**
Data-supported (DS)							
FICO	ElliCE	$\mathbf{1.00 \pm 0.00}$	3.53 ± 0.17	$\mathbf{1.00 \pm 0.00}$	4.91 ± 0.22	$\mathbf{1.00 \pm 0.00}$	5.06 ± 0.29
	DeltaRob	$\mathbf{1.00 \pm 0.00}$	4.00 ± 0.10	$\mathbf{1.00 \pm 0.00}$	5.67 ± 0.58	0.96 ± 0.07	5.70 ± 0.72
	T:Rex	0.83 ± 0.08	3.12 ± 0.07	0.01 ± 0.00	3.07 ± 0.11	0.00 ± 0.00	2.77 ± 0.19
German	ElliCE	$\mathbf{1.00 \pm 0.00}$	3.48 ± 0.10	$\mathbf{1.00 \pm 0.00}$	4.32 ± 0.31	$\mathbf{1.00 \pm 0.00}$	4.00 ± 0.24
	DeltaRob	0.98 ± 0.01	3.45 ± 0.06	0.99 ± 0.02	4.00 ± 0.15	$\mathbf{1.00 \pm 0.00}$	3.99 ± 0.22
	T:Rex	0.99 ± 0.01	3.47 ± 0.04	0.97 ± 0.02	4.03 ± 0.20	0.99 ± 0.01	4.23 ± 0.24
Continuous (CNT)							
Diabetes	ElliCE	$\mathbf{0.98 \pm 0.01}$	2.15 ± 0.39	$\mathbf{0.99 \pm 0.02}$	3.05 ± 0.34	$\mathbf{0.98 \pm 0.02}$	3.22 ± 0.40
	PROP	0.48 ± 0.48	2.01 ± 0.05	0.19 ± 0.28	2.01 ± 0.05	0.08 ± 0.19	2.01 ± 0.05
	ROAR	0.86 ± 0.11	1.86 ± 0.24	0.40 ± 0.28	1.86 ± 0.24	0.31 ± 0.26	1.86 ± 0.24
	T:Rex	0.94 ± 0.03	2.47 ± 0.86	0.90 ± 0.08	4.18 ± 0.36	0.94 ± 0.04	4.18 ± 0.36

Table 2: Runtime performance and speedups for data-supported CE for MLP.

	Absolute (seconds)			Relative (speedup)	
Dataset	**ElliCE**	**T:Rex**	**Delta Rob**	**Over T:Rex**	**Over Delta Rob**
FICO	1.792 ± 0.123	7.006 ± 0.058	242.035 ± 1.161	$3.91\times$	$135.04\times$
COMPAS	0.526 ± 0.011	3.534 ± 0.128	360.480 ± 6.701	$6.72\times$	$685.34\times$
Australian	0.057 ± 0.011	0.281 ± 0.006	2.783 ± 0.032	$4.92\times$	$48.64\times$
Diabetes	0.053 ± 0.001	0.296 ± 0.006	1.922 ± 0.032	$5.60\times$	$36.33\times$
German	0.101 ± 0.001	0.432 ± 0.013	9.905 ± 0.068	$4.27\times$	$97.88\times$

6.1 ElliCE Generates Robust Counterfactuals

Figure 1 illustrates the relationship between the evaluators' multiplicity level $\varepsilon_{\text{target}}$ and the achieved robustness for the baselines. We report results for both linear models and MLPs for data-supported and continuous methods. Across different settings, we observe that ElliCE consistently produces more robust counterfactuals than baselines. Notably, ElliCE's counterfactuals generally do not exhibit a decrease in robustness as $\varepsilon_{\text{target}}$ increases, demonstrating stability under different levels of target multiplicity. This robustness, however, can sometimes come with a greater distance from the original instance (i.e., longer CEs), a trade-off that we saw in Section 5 and report in Table 1. For the MLP setting, our empirical results in Figure 1 and Table 1 suggest that ElliCE's ellipsoidal approximation offers good flexibility, allowing it to adapt to the underlying loss function's shape.

6.2 ElliCE is Efficient

Tables 2, 5 and 6 clearly demonstrates ElliCE's advantage in computational efficiency. Our method is consistently one to three orders of magnitude faster than baselines. Moreover, the runtimes of both T:Rex and Delta Robustness tend to grow substantially with the dataset size. In contrast, ElliCE remains lightweight and exhibits better scalability. Across all datasets tested, ElliCE's absolute runtimes for generating a counterfactual remain under two seconds. This efficiency comes from a closed-form solution for the inner optimization problem (Theorem 1). The primary preprocessing cost involves computing and inverting the Hessian matrix H, requiring $O(np^2)$ for computation and $O(p^3)$ for inversion, performed once per model (where n is the training set size and p is the parameter dimension). Per-instance counterfactual generation then requires only $O(p^2)$ operations.

6.3 Sensitivity Analysis

Figure 2 (a,b) shows an empirical sensitivity analysis of ElliCE's robustness with respect to its internal Rashomon parameter ϵ. The plots show how the achieved robustness (evaluated against the Random Retrain and Ellipsoidal Rashomon set evaluators, respectively) varies as ElliCE's internal ϵ changes. These results illustrate that ElliCE can achieve high levels of robustness even for rel-

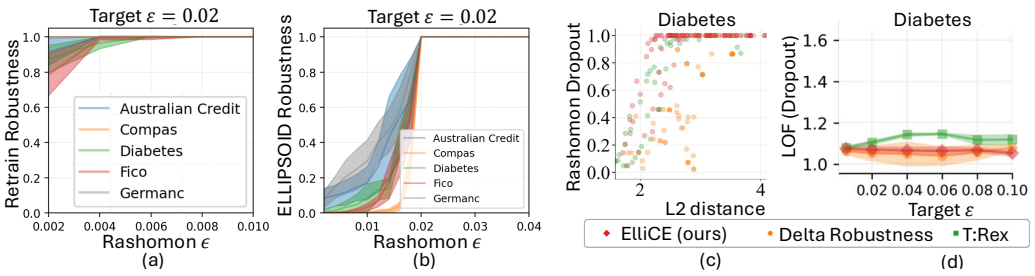

Figure 2: (a,b) Sensitivity of ElliCE's robustness (y-axis) to its internal ϵ hyperparameter (x-axis). Robustness is evaluated against Random Retrain (a) and an Ellipsoidal Rashomon set approximation defined with a fixed $\varepsilon_{\text{target}}$ (b). (c, d) Robustness vs. ℓ_2 proximity trade-off (c) and plausibility (d) of counterfactuals generated by ElliCE and baselines on Diabetes dataset.

atively small values of its internal ϵ when evaluated against the Retrain ensemble. For the middle plot (Ellipsoidal evaluator), while initial robustness may be lower for smaller internal ϵ values, the performance increases sharply, as ϵ approaches the targeted robustness level.

6.4 Robustness-Proximity Trade-off and Plausibility

Figure 2(c) illustrates the inherent trade-off between robustness and proximity for CEs generated by ElliCE, supporting our discussion in Section 5. While the trade-off occurs for all baselines, ElliCE achieves the highest robustness at a given length level. Understanding this trade-off is key to selecting counterfactuals that balance reliability under model shifts with practical user actionability. ElliCE provides a mechanism to navigate this by allowing control over its Rashomon parameter. We also observed good plausibility across all baselines and datasets, as supported by Figure 2(d) and 8. All LOF values tend to be close to 1, thus the generated counterfactuals lie on the data manifold.

6.5 Actionability

To ensure that generated recourse remains realistic and feasible, we incorporate actionability constraints that specify which features can change and within what ranges. ElliCE supports restrictions on features, including immutable features (e.g., age, citizenship) as well as range and direction constraints such as income or loan duration. It also allows for sparse counterfactuals by adding an optional penalty on the number of modified features. For instance, before applying actionability, one robust counterfactual on the German Credit dataset suggested changing the applicant's age, an immutable feature. After enforcing immutability and sparsity constraints, ElliCE instead adjusted the credit amount and credit length, reducing both and thus lowering the predicted credit risk, which is reasonable in the lending context. Further details are provided in Appendix D.

7 Conclusions, Implications and Limitations

Standard algorithmic recourse is fragile. A recommendation given to a user today may become invalid tomorrow if the underlying model is retrained or replaced – a common scenario under the Rashomon effect. This paper addressed this reliability gap by introducing ElliCE, a framework that provides recourse with provable robustness guarantees. ElliCE approximates the set of near-optimal models with an ellipsoid and computes counterfactuals that remain valid across this approximated Rashomon set. A strength of ElliCE is its support for actionability. Users can specify immutable features, range or direction constraints, and optional sparsity penalties, ensuring that the resulting recourse is both robust and realistic. This flexibility might help prevent impractical or unethical recommendations and gives users greater control over actions. While robustness alone does not ensure fairness, user-specified actionability constraints can help to ensure that counterfactuals remain feasible and ethically sound. A comprehensive fairness analysis remains an important direction for future work. The ellipsoidal approximation, while efficient, is a simplification of the true Rashomon set, and for neural networks our analysis currently captures local rather than global model multiplicity. Despite these limitations, ElliCE provides a practical and theoretically grounded tool for robust and actionable recourse, providing stable and trustworthy advice.

Acknowledgments

We thank the RAI for Ukraine program, led by the Center for Responsible AI at New York University in collaboration with Ukrainian Catholic University in Lviv, for supporting Bohdan's and Iryna's participation in this research.

Code Availability

Implementations of ElliCE are available at `https://github.com/BogdanTurbal/ElliCE_EXPERIMENTS`.

References

[1] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:4699–4711, 2021.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, May 2016. [Newspaper article].

[3] Samy Baladram. Iris dataset extended, 2023.

[4] Ainhize Barrainkua, Giovanni De Toni, Jose Antonio Lozano, and Novi Quadrianto. Who pays for fairness? rethinking recourse under social burden. *arXiv preprint arXiv:2509.04128*, 2025.

[5] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022.

[6] Zachery Boner, Harry Chen, Lesia Semenova, Ronald Parr, and Cynthia Rudin. Using noise to infer aspects of simplicity without learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[7] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[8] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

[9] Dieter Brughmans, Pieter Leyman, and David Martens. NICE: An algorithm for nearest instance counterfactual explanations. *CoRR*, abs/2104.07411, 2021. doi: 10.48550/arXiv.2104.07411.

[10] Mustafa Cavus, Jan N van Rijn, and Przemysław Biecek. Beyond the single-best model: Rashomon partial dependence profile for trustworthy explanations in automl. In *International Conference on Discovery Science*, pages 445–459. Springer, 2025.

[11] Marina Ceccon, Alessandro Fabris, Goran Radanović, Asia J Biega, and Gian Antonio Susto. Reinforcement learning for durable algorithmic recourse. *arXiv preprint arXiv:2509.22102*, 2025.

[12] Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T.

[13] Gordon Dai, Pavan Ravishankar, Rachel Yuan, Daniel B. Neill, and Emily Black. Be intentional about fairness!: Fairness, size, and multiplicity in the rashomon set. *CoRR*, abs/2501.15634, 2025. doi: 10.48550/arXiv.2501.15634.

[14] Jiayun Dong and Cynthia Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020.

[15] Jon Donnelly, Srikar Katta, Cynthia Rudin, and Edward P Browne. The rashomon importance distribution: Getting RID of unstable, single model-based variable importance. In *Neural Information Processing Systems (NeurIPS)*, 2023.

[16] Jon Donnelly, Zhicheng Guo, Alina Jade Barnett, Hayden McTavish, Chaofan Chen, and Cynthia Rudin. Rashomon sets for prototypical-part networks: Editing interpretable models in real-time. *arXiv preprint arXiv:2503.01087*, 2025.

[17] Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. Robust counterfactual explanations for tree-based ensembles. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5742–5756. PMLR, Jul 2022.

[18] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 2020.

[19] Fair Isaac Corporation (FICO). FICO explainable machine learning challenge: Home equity line of credit (heloc) dataset, 2018.

[20] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[21] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[22] Alexandre Forel, Axel Parmentier, and Thibaut Vidal. Don't explain noise: Robust counterfactuals for randomized ensembles. *CoRR*, abs/2205.14116, 2022. doi: 10.48550/arXiv.2205.14116.

[23] Prakhar Ganesh, Afaf Taik, and Golnoosh Farnadi. Systemizing multiplicity: The curious case of arbitrariness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1032–1048, 2025.

[24] Prateek Garg, Lokesh Nagalapatti, and Sunita Sarawagi. From search to sampling: Generative models for robust algorithmic recourse. *arXiv preprint arXiv:2505.07351*, 2025.

[25] Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. Interpretable credit application predictions with counterfactual explanations. *CoRR*, abs/1811.05245, 2018. doi: 10.48550/arXiv.1811.05245.

[26] Abirami Gunasekaran, Pritesh Mistry, and Minsi Chen. Which explanation should be selected: A method agnostic model class reliance explanation for model and explanation multiplicity. *SN Computer Science*, 5:503, 2024. doi: 10.1007/s42979-024-02810-8.

[27] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. *CoRR*, abs/2305.11997, 2023. doi: 10.48550/arXiv.2305.11997. International Conference on Machine Learning (ICML), 2023.

[28] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[29] Hsiang Hsu and Flavio Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. In *Neural Information Processing Systems (NeurIPS)*, volume 35, pages 28988–29000, 2022.

[30] Hsiang Hsu and Flavio du Pin Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. *CoRR*, abs/2206.01295, 2022. doi: 10.48550/arXiv.2206.01295. NeurIPS 2022 camera-ready version.

[31] Hsiang Hsu, Ivan Brugere, Shubham Sharma, Freddy Lecue, and Richard Chen. Rashomongb: Analyzing the rashomon effect and mitigating predictive multiplicity in gradient boosting. *Advances in Neural Information Processing Systems*, 37:121265–121303, 2024.

[32] Hsiang Hsu, Guihong Li, Shaohan Hu, and Chun-Fu Chen. Dropout-based rashomon set exploration for efficient predictive multiplicity estimation. In *The Twelfth International Conference on Learning Representations*, 2024.

[33] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Formalising the robustness of counterfactual explanations for neural networks. *CoRR*, abs/2208.14878, 2022. doi: 10.48550/arXiv.2208.14878.

[34] Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. Provably robust and plausible counterfactual explanations for neural networks via robust optimisation. *CoRR*, abs/2309.12545, 2023. doi: 10.48550/arXiv.2309.12545. Accepted at ACML 2023, camera-ready version.

[35] Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. Provably robust and plausible counterfactual explanations for neural networks via robust optimisation. In Berrin Yanıkoğlu and Wray Buntine, editors, *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pages 582–597. PMLR, 11–14 Nov 2024.

[36] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Interval abstractions for robust counterfactual explanations. *CoRR*, abs/2404.13736, 2024. doi: 10.48550/arXiv.2404.13736. Published in Artificial Intelligence Journal.

[37] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Recourse under model multiplicity via argumentative ensembling. In Natasha Alechina, Virginia Dignum, Mehdi Dastani, and Juan S. Sichman, editors, *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)*. International Foundation for Autonomous Agents and Multiagent Systems, 2024.

[38] Junqi Jiang, Luca Marzari, Aaryan Purohit, and Francesco Leofante. Robustx: Robust counterfactual explanations made easy. *arXiv preprint arXiv:2502.13751*, 2025.

[39] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *CoRR*, abs/1907.09615, 2019. doi: 10.48550/arXiv.1907.09615.

[40] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 895–905. PMLR, Aug 2020.

[41] Keita Kinjo. Robust counterfactual explanations under model multiplicity using multi-objective optimization. *CoRR*, abs/2501.05795, 2025. doi: 10.48550/arXiv.2501.05795. Submitted to ICLR 2025.

[42] Gunnar König, Hidde Fokkema, Timo Freiesleben, Celestine Mendler-Dünner, and Ulrike von Luxburg. Performative validity of recourse explanations. *arXiv preprint arXiv:2506.15366*, 2025.

[43] Alejandro Kuratomi, Zed Lee, Panayiotis Tsaparas, Evaggelia Pitoura, Tony Lindgren, Guilherme Dinis Junior, and Panagiotis Papapetrou. Subgroup fairness based on shared counterfactuals. *Knowledge and Information Systems*, pages 1–39, 2025.

[44] Phone Kyaw, Kshitij Kayastha, and Shahin Jabbari. Optimal robust recourse with l^p - bounded model change. *arXiv preprint arXiv:2509.21293*, 2025.

[45] Lucas Langlade, Julien Ferry, Gabriel Laberge, and Thibaut Vidal. Fairness and sparsity within rashomon sets: Enumeration-free exploration and characterization. *arXiv preprint arXiv:2502.05286*, 2025.

[46] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *CoRR*, abs/1712.08443, 2017. doi: 10.48550/arXiv.1712.08443.

[47] Volker Lohweg. Banknote Authentication. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C55P57.

[48] Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6765–6774, 2020.

[49] Anna P Meyer, Yea-Seul Kim, Aws Albarghouthi, and Loris D'Antoni. Perceptions of the fairness impacts of multiplicity in machine learning. *arXiv preprint arXiv:2409.12332*, 2024.

[50] Kiarash Mohammadi, Amir-Hossein Karimi, Gilles Barthe, and Isabel Valera. Scaling guarantees for nearest counterfactual explanations. *CoRR*, abs/2010.04965, 2020. doi: 10.48550/arXiv.2010.04965.

[51] Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. An empirical evaluation of the rashomon effect in explainable machine learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 462–478. Springer, 2023.

[52] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: Generating counterfactuals for real-time recourse and interpretability using residual gans. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1488–1497. PMLR, Aug 2022.

[53] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020 (WWW '20)*, pages 3126–3132. ACM / IW3C2, 2020. doi: 10.1145/3366423.3380087.

[54] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and actionable counterfactual explanations. *CoRR*, abs/1909.09369, 2019. doi: 10.48550/arXiv.1909.09369. Presented at AAAI/ACM Conference on AI, Ethics, and Society 2020.

[55] Ross Quinlan. Statlog (Australian Credit Approval). UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C59012.

[56] Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Amazing things come from having many good models. *arXiv preprint arXiv:2407.04846*, 2024.

[57] Chris Russell. Efficient search for diverse coherent explanations. *CoRR*, abs/1901.04909, 2019. doi: 10.48550/arXiv.1901.04909.

[58] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, 2022.

[59] Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. A path to simpler models starts with noise. In *Advances in Neural Information Processing Systems*, 2023.

[60] John W. Smith, William A. Everhart, William C. Dickson, William C. Knowler, and Richard S. Johannes. Pima Indians Diabetes Database, 1988.

[61] Yiyang Sun, Zhi Chen, Vittorio Orlandi, Tong Wang, and Cynthia Rudin. Sparse and faithful explanations without sparse models. *CoRR*, abs/2402.09702, 2024. doi: 10.48550/arXiv.2402.09702. Accepted in AISTATS 2024.

[62] Athanasios Tsanas and Max Little. Parkinsons Telemonitoring. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C5ZS3N.

[63] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. *CoRR*, abs/2102.13620, 2021. doi: 10.48550/arXiv.2102.13620. Last revised July 13, 2021.

[64] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. *arXiv preprint arXiv:2102.13620*, 2021.

[65] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. *CoRR*, abs/1809.06514, 2018. doi: 10.48550/arXiv.1809.06514. Extended version. ACM Conference on Fairness, Accountability and Transparency (FAT* 2019).

[66] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017. doi: 10.48550/arXiv.1711.00399.

[67] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole Rashomon set of sparse decision trees. In *Neural Information Processing Systems (NeurIPS)*, volume 35, pages 14071–14084, 2022.

[68] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[69] Jayanth Yetukuri, Ian Hardy, Yevgeniy Vorobeychik, Berk Ustun, and Yang Liu. Providing fair recourse over plausible groups. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21753–21760, Mar 2024. doi: 10.1609/aaai.v38i19.30175.

[70] Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. Exploring and interacting with the set of good sparse generalized additive models. In *Advances in Neural Information Processing Systems*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are consistent with the paper's scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitation in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the theorems in the main paper and proofs in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental setups are detailed in the Experimental Section and the Appendix. The code is available in the supplement.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets used are publicly available (Australian Credit, COMPAS, Diabetes, FICO, German Credit). The paper provides sufficient algorithmic details and experimental settings to reproduce results. Code is available in the supplement.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See Section 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report mean and standard deviation over multiple runs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have mentioned the broader impact in the introduction and conclusion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper doesn't release models that have the potential to cause harm.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use open access datasets and baselines and cite the sources of all the datasets and baselines we used in the paper.835

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the code for this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowd-sourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowd-sourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We used LLM for editing and improving the clarity of wording.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.