ABSTRACT

Title of Dissertation: TRUSTWORTHY AND EXPLAINABLE MACHINE LEARNING USING INFORMATION-THEORETIC METHODS

Faisal Adamu Hamman
Doctor of Philosophy, 2025

Dissertation Directed by: Professor Sanghamitra Dutta
Department of Electrical and Computer Engineering

Machine learning is being increasingly deployed in high-stakes domains such as finance, healthcare, and education, in ways that profoundly impact people's lives. However, the decision-making process of these complex black-box models is difficult to understand for human stakeholders such as auditors, institutions, and end-users, raising questions about their adoption, accountability, and trust. Regulatory and ethical standards increasingly advocate for reliable and trustworthy explanations for automated decision-making, particularly in the case of adverse actions and denials. Trustworthy adoption of machine learning requires more systematic and mathematically rigorous approaches for explainability, as many existing methods rely on heuristics that may lack consistency. *This dissertation seeks to address emergent challenges in explainable and trustworthy machine learning by developing novel mathematical frameworks deep-rooted in information-theoretic methods.*

An emerging problem in explainability is the problem of robust counterfactual explanations. Counterfactual explanations (CFEs) guide toward changing the outcome of a model with minimum input perturbation, e.g., increase your income by 10K to qualify for a loan. However, such CFEs can

often become invalid if the machine learning model is updated even slightly. Models are in fact updated quite frequently, leading to the phenomenon of model multiplicity (also known as the Rashomon effect), where multiple models with comparable performance make conflicting predictions on the same input. Such variability can cause previously issued CFEs to become invalid, undermining user trust and the reliability of algorithmic recourse. To address this challenge, we propose a measure called Stability, that captures the robustness of CFEs under natural model updates. We develop practical algorithms to generate robust CFEs for neural networks along with theoretical guarantees, providing a principled foundation for reliable algorithmic recourse in evolving machine learning systems.

Going beyond neural networks, we observe that Tabular large language models (LLMs) are also affected by model multiplicity after fine-tuning. Such variability in predictions can raise concerns about the reliability of Tabular LLMs even as they generate interest in critical domains such as finance for classification with limited labeled data. Interestingly, our stability measure helps quantify the consistency of individual predictions for Tabular LLMs without expensive model retraining and ensembling. Our measure quantifies a prediction's consistency by analyzing (sampling) the model's local behavior around the input in its embedding space. We provide probabilistic guarantees on prediction consistency across a broad class of fine-tuned models, along with experiments on Tabular LLMs.

Expanding our investigation of LLM consistency, we move beyond the classification setting and explore generative LLMs. In retrieval augmented generation (RAG) systems, users might expect that paraphrased or reworded queries will yield outputs that convey the same underlying information. However, existing RAG pipelines often show variability in both the retriever and the generator, undermining reliability in high-stakes applications. To address this challenge, we propose a reinforcement-learning-based approach that improves consistency through group similarity rewards computed over paraphrased input sets. Our training strategy yields Con-RAG, a reliable RAG system that improves

both consistency and accuracy across several question-answering (QA) benchmarks.

In addition to improving inference-time consistency and robustness, we explore how CFEs can enhance training-time efficiency in LLMs. We propose CFE-infused Distillation (CoD), a new framework to distill large teacher models into smaller students using few-shot task-specific data by systematically infusing training data with CFE examples. We provide both statistical and geometric guarantees motivating this approach, and show empirically that CoD significantly outperforms standard distillation baselines in few-shot regimes, thus connecting explainability with model compression.

Finally, we turn to another dimension of trust - algorithmic fairness - where we explain fairness trade-offs using information-theory. We reveal how local fairness (within each client) and global fairness (across clients) interact under data heterogeneity in distributed and federated learning by using a new tool in information theory called Partial Information Decomposition (PID). We further unify classical group fairness notions, e.g., statistical parity, equalized odds, and predictive parity using the same PID framework, offering a granular understanding of their overlaps and impossibilities.

This dissertation lays the foundational principles for data scientists and policymakers toward trustworthy AI adoption, reducing reputational and regulatory risks, and fostering user acceptance.

TRUSTWORTHY AND EXPLAINABLE MACHINE LEARNING USING INFORMATION-THEORETIC METHODS

by

Faisal Adamu Hamman

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
November, 2025

Advisory Committee:
Professor Sanghamitra Dutta, Chair/Advisor
Professor Sennur Ulukus
Professor Furong Huang
Professor Dinesh Manocha
Professor Tudor Dumitras

Acknowledgments

"It takes a village to raise a child." — African Proverb

This timeless proverb has never felt more true to me than through the course of this Ph.D. journey. It speaks to the idea that growth and success are never achieved in isolation, they are built through the support and kindness of a community. In many ways, this dissertation is not just the product of my individual effort, but the reflection of the countless people who have guided, encouraged, and believed in me along the way. It truly took a village to complete this work.

I would like to express my deepest gratitude to my advisor, Prof. *Sanghamitra Dutta*, for her unwavering support, patience, and belief in me. I began this journey with little research experience, and through her mentorship I grew into an independent researcher. I am profoundly grateful for the opportunity to work with her over these years.

I am deeply thankful to my parents, *Adamu Hamman* and *Aisha Hamman*, whose endless support, prayers, and sacrifices have made everything possible. I also thank my siblings, *Aishah*, *Imran*, *Sumaiyah*, and my entire family for their constant love and encouragement.

I sincerely thank my Ph.D. committee members: Prof. *Sennur Ulukus*, Prof. *Furong Huang*, Prof. *Dinesh Manocha*, and Prof. *Tudor Dumitras* for serving on my committee and for their thoughtful feedback on my work. I also wish to thank my undergraduate professors at Işık University, Prof. *Onur Kaya*, who first introduced me to information theory (which later became the core of this dissertation), Prof. *Yorgo İstefanopulos*, who encouraged me to pursue a Ph.D., and Prof. *Ümit Güz* and Prof. *Ramazan Köprü* for their mentorship in early academic years.

I am deeply thankful to my collaborators and mentors at J.P. Morgan, *Sanjay Kariyappa*, from whom I have learned a great deal, *Freddy Lecue*, and *Saumitra Mishra*. Our work together forms

a major part of this dissertation, and I'm grateful for the opportunity to work with them. I am also grateful to my mentors at Capital One, *Chenyang Zhu*, *Anoop Kumar*, *Alfy Samuel*, and *Daben Liu*, for their guidance and mentorship during my internship. I would like to thank my lab mates *Pasan*, *Barproda*, and *Yanjun*, as well as my collaborators *Jiahao Chen*, *Erfaun Noorani*, *Richard Zhang*, and *Ilia Sucholutsky*, for their support and insightful discussions throughout my Ph.D.

To my friends, *Mohamed*, *Sondos*, *Nehal*, *Faizan*, *Ahmad Adel*, *Cherif*, *Ashry*, *Tawfiq*, *Marwan*, *Atwa*, *Ayo*, *Temitayo*, this journey simply wouldn't have been possible without you. Thank you to *Jouri Ghazi* and her family, who have been like family to me here, helping me with countless presentations and posters. I also want to extend my appreciation to *Abdulrahman* and *Bluestream Education*, who have guided and supported me from the very beginning. To the Carless Ones, *Rafsun*, *Abdullah*, *Omar*, *Zavian*, *Sam*, thank you for our unmatched late-night chats. I also thank my undergrad friends, *Salim*, *Abdulwahab*, *Kurfi*, *Munira*, *Yusuf*, *Nadia*, *Israa*, *Ismail*, *Sherif*, *Farouk*, *Bilal*, *Ghenna*, *Sadiaa*, *Fatima*, *Ekin*, *Basak*, *Marium*, *Onur*, *Zainab*, *Yasmin*, *Moh Sarki*, whose support from years ago formed the foundation on which this journey was built. To my Capital Science friends, thank you for your support all these years. To everyone I've played football with, from pickup games to tournaments, it has been my escape, keeping me balanced through the highs and lows of the Ph.D. program.

I am also honored and humbled by the recognition I received from the ECE department at UMD. I thank them for recognizing my work through the ECE Ph.D. Distinguished Dissertation Award, the George Corcoran Memorial Award in Teaching, the Outstanding TA Awards, and the TA Training & Development Fellowship. These honors mean a great deal to me and reflect the incredible academic environment I've been fortunate to be part of.

This Ph.D. is not mine alone. It is the collective product of a community that believed in me, supported me, and inspired me. To each and every one of you—thank you!

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The growing use of machine learning (ML) in high-stakes applications such as finance, healthcare, and education brings significant performance benefits but also raises concerns around transparency, robustness, and trust [4–7]. As automated decisions shape crucial outcomes, stakeholders often struggle to understand the reasons behind their predictions. Regulatory and ethical standards require ML models to offer trustworthy explanations for their decisions. For instance, the U.S. Equal Credit Opportunity Act (ECOA) mandates lenders to explain credit denials [8], while the EU's GDPR upholds the right to explanations, particularly for adverse actions and denials [9]. These explanations go beyond compliance: they enable individuals to understand, trust, and, if needed, challenge the automated decisions that impact their lives [10, 11]. This motivates an urgent need for a more rigorous mathematical perspective on explainable machine learning, as many existing methods rely on heuristics that may lack consistency [10, 11]. This dissertation proposes to address this need by leveraging information-theoretic methods to meet emergent challenges in explainable ML, providing more reliable and fair insights into model decisions for trustworthy adoption of AI. The dissertation statement is as follows: *To address emergent challenges in explainable and trustworthy machine learning by developing novel mathematical frameworks deep-rooted in information-theoretic methods.*

In this dissertation, we start with an emergent problem in explainability – the problem of robust algorithmic recourse under model multiplicity (see Chapter 2). Model multiplicity, also known

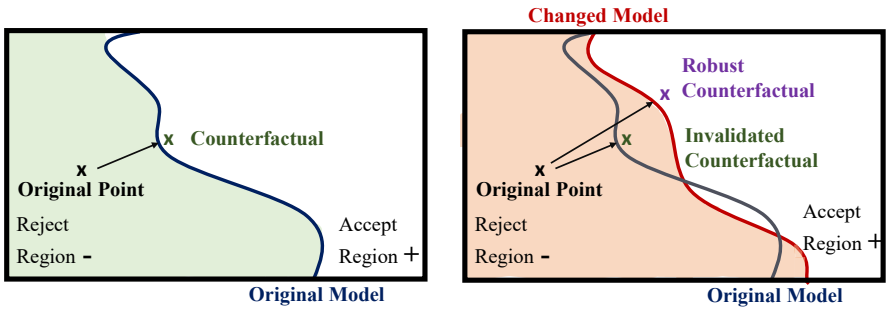


Figure 1.1: *(left)* Original model's decision boundary: Given a point in the rejected region, a counterfactual explanation (counterfactual) typically refers to the closest point on the accepted side. Counterfactual explanations provide guidance on actions for recourse. *(right)* A slightly updated model for which the initial counterfactual becomes invalid, leading to reputational risks and user distrust. A robust counterfactual should remain valid even when the model changes while maintaining proximity to the original data point.

as the Rashomon effect or predictive multiplicity [12–15], is a phenomenon where multiple, equally well-performing models can yield conflicting predictions for the same input due to random variations in the training process, such as seed and random weight initialization. Multiplicity can significantly affect the robustness of counterfactual explanations which are essentially institutional suggestions for flipping a model outcome with minimal effort, e.g., increase income by 10K to qualify for a loan (see Figure 1.1 for illustration). We introduce a new measure called Stability to quantify the robustness of counterfactuals to model multiplicity, along with theoretical guarantees and practical algorithms to find robust counterfactuals for neural networks. Going beyond neural networks, we study the robustness and consistency of predictions under model multiplicity in large language models (LLMs) fine-tuned for tabular data, a setup that is increasingly generating interest in high-stakes applications such as finance for classification with limited labeled data (see Chapter 3). Going beyond classification tasks, we next study the output consistency of generative LLMs under semantically equivalent inputs. We focus on achieving *information consistency*, i.e., paraphrased or semantically similar queries should produce consistent outputs that preserve the underlying informational content (see Chapter 4). In ad-

dition to improving inference-time consistency, we also explore the use of explanations for improving training-time efficiency. Specifically, we propose the use of counterfactual explanations as tools for efficient few-shot knowledge distillation of LLMs (see Chapter 5). We leverage CFEs as informative probes that help student models better approximate teacher decision boundaries under data-scarce settings, supported by theoretical motivations and empirical gains. Finally, we turn to another dimension of trust - algorithmic fairness - where we investigate the fundamental trade-offs among group fairness notions in both decentralized (see Chapter 6) and centralized settings (see Chapter 7). *In essence, this dissertation lay the foundational guiding principles for data scientists and policymakers toward trustworthy AI adoption, reducing reputational and regulatory risks, and fostering user acceptance.*

The following sections provide a detailed overview of each chapter and its specific contributions.

Chapter 2: Robust Counterfactual Explanations Under Model Multiplicity. Counterfactual explanations are changes to input features that would alter the prediction of an ML model, providing actionable insights into how an individual could achieve a desired outcome (algorithmic recourse). There is an emerging interest in generating robust algorithmic recourse that would remain valid if the model is updated or changed even slightly (model multiplicity). Towards finding robust counterfactual explanations, existing literature often assumes that the original model m and the new model M are bounded in the parameter space, i.e., $\|\text{Params}(M) - \text{Params}(m)\| < \Delta$. However, we observe that models can often change significantly in the parameter space with little to no change in their predictions or accuracy on the given dataset. In this work, we introduce a mathematical abstraction, termed *naturally-occurring* model change, which allows for arbitrary changes in the parameter space such that the change in predictions on points that lie on the data manifold is limited. We propose a measure – that we call *Stability* – to quantify the robustness of counterfactuals to potential model changes for differentiable models, e.g., neural networks. Our main contribution is to show that counterfactuals

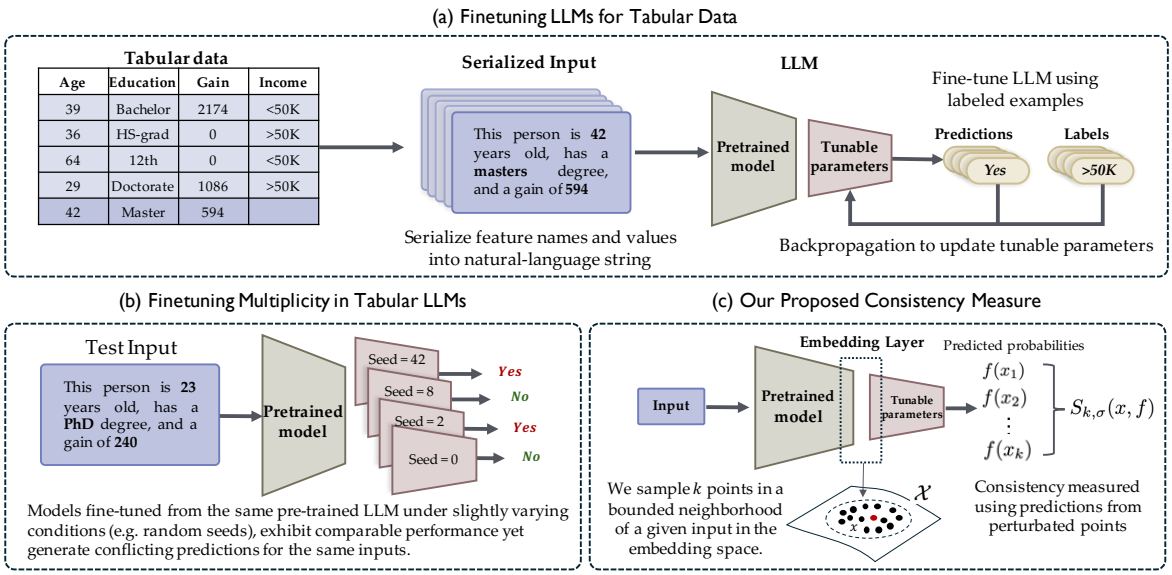


Figure 1.2: (*a*) illustrates the process of fine-tuning LLMs for Tabular data using few labeled examples. (*b*) demonstrates the concept of finetuning multiplicity. (*c*) introduces our proposed local stability measure used to quantify the consistency of individual predictions without requiring the retraining of multiple models.

with sufficiently high value of *Stability* as defined by our measure will remain valid after potential "naturally-occurring" model changes with high probability (leveraging concentration bounds for Lipschitz functions of independent Gaussians). Since our quantification depends on the local Lipschitz constant around a data point which is not always available, we also examine estimators of our proposed measure and derive a fundamental lower bound on the sample size required to have a precise estimate. We explore methods of using stability measures to generate robust counterfactuals that are close, realistic, and remain valid after potential model changes.

Chapter 3: Robust Predictions in Tabular LLMs Under Model Multiplicity. LLMs are increasingly used in high-stakes tabular classification tasks, particularly due to their commendable performance in scenarios with limited training data [1]. However, we find that fine-tuning LLMs on tabular data can lead to significant multiplicity (see Figure 1.2). This raises critical concerns about the robustness and reliability of Tabular LLMs when deployed in critical applications. This chapter formalizes

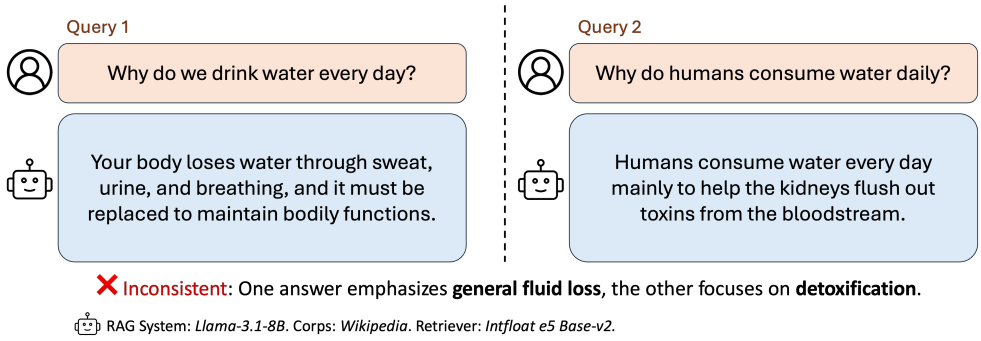


Figure 1.3: Information Consistency in RAG Systems. Two semantically equivalent queries lead to different outputs from a RAG system, despite both responses being factually correct. Such variation may be acceptable in many applications, but in certain domains information consistency across similar inputs may be required to ensure reliability, user trust, and compliance.

the unique challenge of fine-tuning multiplicity in Tabular LLMs and leverages our stability measure to quantify the robustness of individual predictions without expensive LLM retraining. Our measure quantifies a prediction's robustness by analyzing (sampling) the model's local behavior around the input in the embedding space. Interestingly, we show that sampling in the local neighborhood can be leveraged to provide probabilistic robustness guarantees against a broad class of equally well-performing fine-tuned models. By leveraging Bernstein's Inequality, we show that predictions with sufficiently high robustness (as defined by our measure) will remain consistent with high probability. We also provide empirical evaluations on real-world datasets to support our theoretical results.

Chapter 4: Information Consistency in Retrieval-Augmented Generation Systems. To expand our investigation of LLM consistency, we move beyond classification settings and explore generative LLMs. In retrieval-augmented generation (RAG) systems, users expect that paraphrased or reworded queries will yield outputs that convey the same underlying information (see Figure 1.3). However, existing RAG pipelines often violate this expectation due to variability in both the retriever and the generator (LLM), undermining reliability in high-stakes applications. We formalize this property as *information consistency*, defined as the requirement that semantically equivalent inputs produce out-

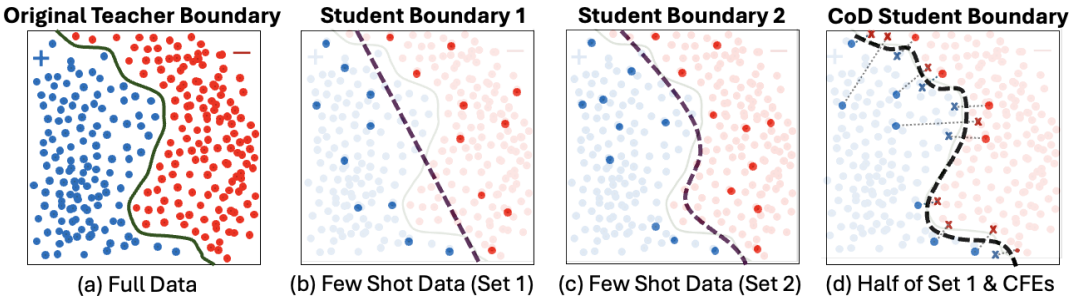


Figure 1.4: Intuition behind Counterfactual Explanations for Knowledge Distillation:*(a)* Teacher trained on the full dataset with true decision boundary. *(b–c)* With few-shot supervision, many classifiers can fit the sparse points; the resulting student boundaries (dashed lines) can vary and do not always align with the teacher's boundary (unfaithful distillation). *(d)* Pairing each point with its CFE (×, linked to originals) during distillation makes the student match the teacher's soft predictions at these points. CFEs act as boundary-near pegs that clamp the student to the teacher's decision surface, producing a more faithful distillation even under few-shot budgets.

puts preserving consistent informational content. To analyze and improve this property, we introduce a principled evaluation framework that decomposes RAG consistency into retriever-level, generator-level, and end-to-end components, enabling a fine-grained understanding of inconsistency sources. To improve consistency, we propose Paraphrased Set Group Relative Policy Optimization (PS-GRPO), an RL method that aligns outputs across paraphrased inputs by assigning group similarity rewards. This framework yields our **Con**sistent **RAG** (Con-RAG) system, which produces outputs that remain stable across semantically equivalent queries and robust to retrieval-induced variability. Empirical evaluations across benchmarks datasets demonstrate that Con-RAG achieves significant improvements in both consistency and accuracy over strong baselines, even without ground-truth supervision.

Chapter 5: Few-Shot Knowledge Distillation of LLMs with Counterfactual Explanations. In addition to improving inference-time consistency and robustness, we explore how CFEs can enhance training-time efficiency in LLMs. We leverage CFE as a tool for efficient learning, using them as *informative probes* for few-shot knowledge distillation of LLMs (see Figure 1.4). Knowledge distillation aims to transfer capabilities from a high-capacity teacher model to a smaller, resource-efficient student,

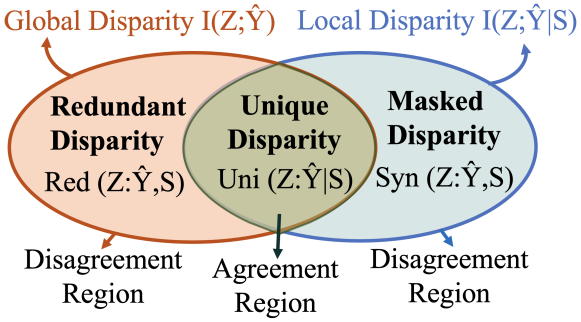


Figure 1.5: Partial Information Decomposition (PID) is used to identify three distinct types of disparity that contribute to global and local disparity in FL: Unique, Redundant, and Masked Disparity. Venn diagram highlights agreement and disagreement regions of global and local fairness.

but its effectiveness often depends on the availability of large task-specific datasets. In many practical settings, such data is limited, leading to suboptimal generalization in few-shot regimes. To address this challenge, we propose **Co**unterfactual-explanation-infused **D**istillation (CoD), a framework that systematically infuses CFEs into the distillation training data. By enriching the few-shot training set with CFEs, CoD enables the student model to better approximate the teacher's decision boundary with significantly fewer samples (see Figure 1.4 for intuition). We provide both statistical and geometric motivations for our approach, and demonstrate empirically that CoD consistently outperforms standard baselines in few-shot regimes. This work bridges explainability and model compression, showing how explanations can serve as learning signals for more data-efficient knowledge transfer.

Chapter 6: Explaining Group Fairness Trade-offs and Impossibilities in Distributed Environments. To further advance the goal of trustworthy ML, we turn to the dimension of fairness. This chapter presents an information-theoretic perspective to explaining group fairness trade-offs in federated learning (see Figure 1.5). In federated learning (FL), multiple clients collaboratively train a global model while keeping their data local, leading to data heterogeneity and uneven representation across certain groups. Recent works focus on finding models that are fair when evaluated on the entire dataset across all clients, a concept known as global fairness. E.g., several banks may decide to engage in FL

to train a model that will determine loan qualifications without exchanging data among them. A globally fair model does not discriminate against any group when evaluated on the entire dataset across all the banks. On the other hand, local fairness considers the disparity of the model at each client (when evaluated on a client's local dataset). Local fairness is important as the models are ultimately deployed and used locally. Existing works often focus on either global fairness or local fairness, without always considering their trade-offs. There is a lack of understanding regarding the interplay between global and local fairness in FL, particularly under data heterogeneity, and if and when one implies the other. To address this gap, we leverage a body of work in information theory called partial information decomposition (PID), which first identifies three sources of unfairness in FL, namely, *Unique*, *Redundant*, and *Masked Disparity* (see Figure 1.5 for illustration). This decomposition helps us derive fundamental limits on the trade-off between global and local fairness, highlighting where they agree or disagree. We introduce the *Accuracy and Global-Local Fairness Optimality Problem* (AGLFOP), a convex optimization that defines the theoretical limits of accuracy and fairness trade-offs, identifying the best possible performance any FL strategy can attain given a dataset and client distribution.

Chapter 7: Explaining Group Fairness Trade-offs and Impossibilities in Centralized Environments. Having examined fairness trade-offs in distributed and federated settings, we now turn to the centralized ML, where all data is available in a single location. This chapter introduces a novel information-theoretic perspective on explaining the relationship between prominent group fairness notions in machine learning, namely statistical parity, equalized odds, and predictive parity (see Figure 1.6). It is well known that simultaneous satisfiability of these three fairness notions is usually impossible, motivating practitioners to resort to approximate fairness solutions rather than stringent satisfiability of these definitions. However, a comprehensive analysis of their interrelations, particularly when they are not exactly satisfied, remains largely unexplored. Our main contribution lies in

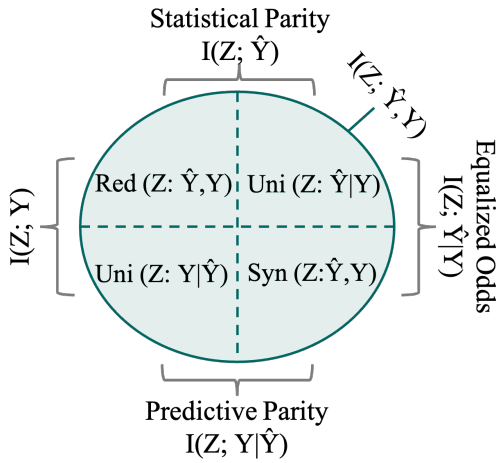


Figure 1.6: Venn diagram showing the exact relationship between the various unfairness measures using PID: A critical observation is that all four PID terms are nonnegative. This enables us to derive several fundamental limits and tradeoffs among the unfairness measures, providing a nuanced understanding of when they agree and disagree.

elucidating an exact relationship between these three measures of (un)fairness by leveraging a body of work in information theory called partial information decomposition (PID). In this work, we leverage PID to identify the granular regions where these three measures of (un)fairness overlap and where they disagree with each other leading to potential tradeoffs.

Bibliographical Overview. The research presented in this dissertation is primarily based on the author's first-authored, peer-reviewed publications. Chapter 2 is based on [16, 17], which study the robustness of counterfactual explanations under model multiplicity. Chapter 3 draws from [18], focusing on prediction consistency in tabular LLMs. Chapter 4 is based on [19], which investigates information consistency in retrieval-augmented generation systems. Chapter 5 builds on [20], exploring counterfactual explanations for few-shot knowledge distillation. Chapter 6 incorporates work from [21, 22], and Chapter 7 from [23], presenting an information-theoretic analysis of group fairness trade-offs. A related work on privacy and algorithmic fairness [24] was conducted but omitted from this dissertation.

Chapter 2: Robust Algorithmic Recourse Under Model Multiplicity

2.1 Introduction

Algorithmic recourse [25–27] have garnered significant interest in various high-stakes applications, such as lending, hiring, etc. Algorithmic recourse aim to guide an applicant on how they can change a model outcome by providing suggestions for improvement. Given an original data-point (e.g., an applicant who is denied a loan), the goal is to try to find a point on the other (desired) side of the decision boundary (a hypothetical applicant who is approved for the loan) which also satisfies several other preferred constraints, such as, (i) proximity to the original point; (ii) changes in as few features as possible; and (iii) conforming to the data manifold. Such a data-point that alters the model decision is widely referred to as a "counterfactual explanation," as illustrated in Fig. 1.1.

However, in several real-world scenarios, such as credit lending, the models have to be updated due to various reasons [28–30], e.g., to retrain on a few additional data points, change the hyper-parameters or seed, or transition to a different model class [31]. Such model changes can often cause the counterfactuals to become invalid because they are typically close to the decision boundary. For instance, suppose the counterfactual explanation suggests an applicant to increase their income by $10K$ to get approved for a loan, and they act upon that, but now, due to updates to the original model, they are still denied by the updated model (see Fig. 1.1).

If recourse becomes invalid due to model updates, this can lead to confusion and distrust in the

use of algorithms in high-stakes applications altogether. Users would typically act on the suggested counterfactuals over a period of time, e.g., increase their income for credit lending, but only to find that it is no longer enough since the model has slightly changed (perhaps due to retraining with a new seed or hyperparameter). This cycle of invalidation and regenerating new counterfactuals can not only be frustrating and time-consuming for users but also potentially hurt an institution's reputation. This motivates our primary question: *How to provide theoretical guarantees on the robustness of counterfactuals to potential model changes?*

Towards addressing this question, in this work, we introduce the abstraction of "naturally-occurring" model change for differentiable models. Our abstraction allows for arbitrary changes in the parameter space such that the change in predictions on points that lie on the data manifold is limited. This abstraction is centered on the inherent need for model explanations to remain robust against variations such as weight initialization or minor adjustments in hyperparameters (changing seed) [28, 29, 32–35]. Another insightful angle is the concept of machine unlearning, particularly in light of regulatory frameworks like the GDPR [9]. The *right to be forgotten* necessitates the removal of an individual's data upon request, potentially leading to model updates. These updates could, in turn, impact the validity of previously issued explanations, thus challenging the balance between the *right to explanation* and the *right to be forgotten* [30].

This abstraction motivates a measure of robustness for counterfactuals that arrives with provable probabilistic guarantees on their validity under naturally-occurring model change. We also introduce the notion of adversarial or targeted model change and provide an impossibility result for such model change. We examine estimators of our proposed measure and derive a fundamental lower bound on the sample size required to have a precise estimate. Next, by leveraging this estimator, we explore methods of using stability measures to generate robust counterfactuals that are close, realistic, and remain valid

after potential model changes. Our experimental results validate our theoretical understanding and illustrate the efficacy of our proposed algorithms. We summarize our contributions here:

Abstraction of "naturally-occurring" model change for differentiable models: Existing literature [28, 29] on robust counterfactuals often assumes that the original model m and the new model M are bounded in the parameter space, i.e., $\|\text{Params}(M) - \text{Params}(m)\| < \Delta$. Building on [32] for tree-based models, we note that models can often change significantly in the parameter space with little to no change on their predictions or accuracy on the given dataset. To capture this, we introduce an abstraction (see Definition 2.5), that we call *naturally-occurring* model change, which instead allows for arbitrary changes in the parameter space such that the change in predictions on points that lie on the data manifold is limited. Our proposed abstraction of naturally-occurring model change also has interesting connections with predictive/model multiplicity, also known as, the Rashomon Effect [12, 13].

We also make a clear distinction between our proposed naturally-occurring and *adversarial* model change. Under the adversarial model change, we provide an impossibility result (Theorem 2.2) that given any counterfactual for a model, one can always design a new model that is quite similar to the original model and that renders that particular counterfactual invalid. However, our focus is on non-targeted model change such as retraining on a few additional data points, changing some hyper-parameters or seed, etc. which is captured in "naturally-occurring" model change (see Definition 2.5).

A measure of robustness with probabilistic guarantees on validity: Next, we propose a novel mathematical measure – that we call Stability – to quantify the robustness of counterfactuals to potential model changes. Stability of a counterfactual $x \in \mathbb{R}^d$ with respect to a model $m(\cdot)$ is given by:

$$R_{k,\sigma^2}(x, m) = \frac{1}{k} \sum_{x_i \in N_{x,k}} \left(m(x_i) - \gamma_x \|x - x_i\| \right),$$

where $N_{x,k}$ is a set of k points in $\mathbb{R}^d$ drawn from the Gaussian distribution $\mathcal{N}(x, \sigma^2 \mathrm{I}_d)$ with I_d being the identity matrix, and γ_x is the local Lipschitz constant of the model $m(\cdot)$ around x (see Definition 2.6).

Our main contribution is to provide a theoretical guarantee (Theorem 2.3) that counterfactuals with a sufficiently high value of Stability (as defined by our measure) will remain valid with high probability after *naturally-occurring* model changes. In Theorem 2.3, we assume a strict upper bound $|\mathbb{E}[Z|M] - \mathbb{E}[Z]| < \epsilon'$, where $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. We generalize this by introducing a probabilistic bound $\Pr(|\mathbb{E}[Z|M] - \mathbb{E}[Z]| > \epsilon') \leq \delta$ (see Corollary 2.1). Further, we characterize this bound δ under the conditions of naturally-occurring model change and specific assumptions about the expected variability in a data point's neighborhood (see Assumption 2.1). Leveraging this characterization, we introduce Lemma 2.3, which serves as the foundation for proving Theorem 2.4. This theorem offers a comprehensive probabilistic guarantee on the validity of counterfactuals with a high value of Stability (as per our measure) on the data manifold. Our results leverage concentration bounds for Lipschitz functions of independent Gaussian random variables (see Lemma 2.2).

Estimators of stability and their properties: Since our proposed measure depends on the local Lipschitz constant which is not known, we examine two practical estimators: (1) The *Stability-Lipschitz estimator* (Definition 2.7) approximates the local Lipschitz constant using $\hat{\gamma}_x = \max_{x_i \in N_{x,k}} \frac{|m(x)-m(x_i)|}{\|x-x_i\|}$. This captures the *worst-case variability* in the model's outputs in the neighborhood of x. We derive a fundamental lower bound on sample size to ensure that the estimator approximates the true stability within an ϵ error (see Theorem 2.5). (2) We introduce the *Stability-Soft estimator* (see Definition 2.8) as a less computationally expensive, albeit less accurate, alternative for estimating stability:

$$\hat{R}_{k,\sigma^2}(x, m) = \frac{1}{k} \sum_{x_i \in N_{x,k}} \left(m(x_i) - |m(x) - m(x_i)|\right).$$

The first term essentially captures the mean value of the model output in a region around it (higher mean is expected to be more robust and reliable). The second term captures the local *average variability* of the model output around it (lower variability is expected to be more reliable). This intuition is in alignment with the results in [32] for tree-based models (see Section 2.4.4).

Generating robust counterfactuals using stability: We explore strategies for using stability measures to generate robust counterfactuals for neural networks. We introduce T-Rex:I (Algorithm 1), which finds robust counterfactuals that are close to the original data point. T-Rex:I can be integrated into any base technique for generating counterfactuals to improve robustness. We also propose T-Rex:NN (Algorithm 2), which generates robust counterfactuals that are data-supported (along the lines of [32] for tree-based models). We propose a hybrid method T-Rex:Hybrid (Algorithm 3) that focuses on finding robust counterfactuals on the data manifold, making them more realistic. The hybrid method employs generative models to learn a latent representation of the data manifold, within which we conduct our search for counterfactuals.

Experimental results: We conduct experiments on several benchmark datasets, namely, HELOC [36], German Credit, Cardiotocography (CTG), Adult [37], and Taiwanese Credit [38] to support our theoretical findings (see Section 2.5). Our experiments show that T-Rex:I can improve robustness for neural networks without significantly increasing the cost, and T-Rex:NN consistently generates counterfactuals that are similar to the data manifold, as measured using the Local Outlier Factor (LOF). The Local Outlier Factor (LOF) is a popular evaluation metric that assesses the relative isolation of a data point within their local neighborhood to identify anomalies (see Definition 2.4).

2.2 Related Works

Counterfactual explanations have seen growing interest in recent years [25, 26, 34]. Regarding their robustness to model changes, [39–41] argue that counterfactuals situated on the data manifold are more likely to be more robust than the closest counterfactuals. Later, [32] demonstrate that generating counterfactuals on the data manifold may not be sufficient for robustness. While the importance of robustness in local explanation methods has been emphasized [42], the problem of specifically generating robust counterfactuals has been less explored, with the notable exceptions of some recent works [28, 29, 32, 33, 43]. In [28], the authors propose an algorithm called ROAR that uses min-max optimization to find the *closest* counterfactuals that are also robust. In [43], the focus is on analytical trade-offs between validity and cost. [33] introduces a method for identifying close and robust counterfactuals based on a framework that utilizes interval neural networks. [29] propose that local Lipschitzness can be leveraged to generate consistent counterfactuals and propose an algorithm called Stable Neighbor Search to generate consistent counterfactuals for neural networks. Our research builds on this perspective and further performs Gaussian sampling around the counterfactual, leading to a novel estimator for which we are also able to provide probabilistic guarantees going beyond the bounded model change assumption. Furthermore, examining all three performance metrics, namely, cost, validity (robustness), and likeness to the data-manifold has received less attention with the notable exception of [32] but they focus only on tree-based models (non-differentiable). Following our conference publication, [44] proposed a robust optimization framework to generate provably robust and plausible counterfactuals for neural networks and proved its soundness, completeness, and convergence. We also refer to [45] for a survey.

We note that [46, 47] propose an alternate perspective of robustness in explanations (called L-

stability in [47]) which is built on similar individuals receiving similar explanations. [48–50] focus on finding counterfactuals that are robust to small input perturbations (noisy counterfactuals). In contrast, our focus is on counterfactuals remaining valid after some changes to the model, and providing theoretical guarantees thereof.

Our work also shares interesting conceptual connections with a body of work on model multiplicity or predictive multiplicity, also known as the Rashomon effect [12–15]. [12] suggested that models can be very different from each other but have almost similar performance on the data manifold. The term predictive multiplicity was suggested by [13] which defined it as the ability of a prediction problem to admit competing models with conflicting predictions. [14] investigates ways to leverage model multiplicity beneficially in model selection processes while simultaneously addressing its concerning implications. [51] offered a framework for measuring predictive multiplicity in classification, introducing measures that encapsulate the variation in risk estimates over the ensemble of competing models. [15] unveiled a novel metric, Rashomon Capacity, for measuring predictive multiplicity in probabilistic classification. Our proposed abstraction of naturally-occurring model change in this work can be viewed as a fresh perspective on model multiplicity that further emphasizes the models that are more likely to occur.

2.3 Preliminaries

Let $m(\cdot) : \mathbb{R}^d \to [0, 1]$ denote the original machine learning model that takes a d-dimensional input value and produces an output probability lying between 0 and 1. The final decision is denoted by $\mathbb{1}(m(x) \geq 0.5)$ where $\mathbb{1}(\cdot)$ denotes the indicator function.

Definition 2.1 ($\gamma-$Lipschitz)**.** *A function $m(\cdot)$ is said to be $\gamma-$Lipschitz if*

$$|m(x) - m(x')| \leq \gamma \|x - x'\| \quad \forall\, x, x' \in \mathbb{R}^d.$$

Here $\|\cdot\|$ denotes the Euclidean norm, i.e., for $u \in \mathbb{R}^d$, we have $\|u\| = \sqrt{u_1^2 + u_2^2 + \ldots + u_d^2}$. In Remark 2.2, we also discuss relaxations to local Lipschitz constants from global Lipschitz constants. We denote the updated or changed model as $M(\cdot) : \mathbb{R}^d \to [0, 1]$ where M is a random entity. We mostly use capital letters to denote random entities, e.g., M, X, etc., and small letters to denote non-random entities, e.g., m, x, γ, n, etc.

Definition 2.2 (Closest Counterfactual $\mathcal{C}_p(x, m)$)**.** *Given $x \in \mathbb{R}^d$ such that $m(x) < 0.5$, its closest counterfactual (in terms of l_p-norm) with respect to the model $m(\cdot)$ is defined as a point $x' \in \mathbb{R}^d$ that minimizes the l_p norm $\|x - x'\|_p$ such that $m(x') \geq 0.5$.*

$$\mathcal{C}_p(x, m) = \arg\min_{x' \in \mathbb{R}^d} \|x - x'\|_p \text{ such that } m(x') \geq 0.5.$$

When one is interested in finding counterfactuals by changing as few features as possible, the l_1 norm is used (enforcing a sparsity constraint). These are called *sparse* counterfactuals [39]. However, such closest counterfactuals often fall too far from the data manifold, resulting in unrealistic and anomalous instances, as noted in [26, 34, 39–41, 52]. This highlights the need for generating counterfactuals that lie on the data manifold.

Definition 2.3 (Closest Data-Manifold Counterfactual $\mathcal{C}_{p,\mathcal{X}}(x, m)$)**.** *Given $x \in \mathbb{R}^d$ such that $m(x) < 0.5$, its closest data-manifold counterfactual $\mathcal{C}_{p,\mathcal{X}}(x, m)$ with respect to the model $m(\cdot)$ and data man-*

ifold $\mathcal{X} \subseteq \mathbb{R}^d$ *is defined as a point* $x' \in \mathcal{X}$ *that minimizes the* l_p *norm* $\|x - x'\|_p$ *such that* $m(x') \geq 0.5$.

$$\mathcal{C}_{p,\mathcal{X}}(x, m) = \arg\min_{x' \in \mathcal{X}} \|x - x'\|_p \text{ such that } m(x') \geq 0.5.$$

In order to assess the similarity or anomalous nature of a point concerning the given dataset $\mathcal{S} \subseteq \mathcal{X}$, various metrics can be employed, e.g., K-nearest neighbors, Mahalanobis distance, Kernel density. These metrics play a crucial role in understanding the quality of counterfactual explanations generated by a model. One metric widely used in literature [32, 39, 40] is the Local Outlier Factor..

Definition 2.4 (Local Outlier Factor (LOF) [53])**.** *For* $x \in \mathcal{S}$*, let* $L_k(x)$ *be its* k*-Nearest Neighbors (k-NN) in* $\mathcal{S}$*. The* k*-reachability distance* rd_k *of* x *with respect to* x' *is defined by* $rd_k(x, x') = \max\{\delta(x, x'), d_k(x')\}$*, where* $d_k(x')$ *is the distance* δ *between* x' *and its* k*-th nearest instance on* $\mathcal{S}$*. The* k*-local reachability density of* x *is defined by* $lrd_k(x) = |L_k(x)|(\sum_{x' \in L_k(x)} rd_k(x, x'))^{-1}$*. Then, the k-LOF of* x *on* $\mathcal{S}$ *is defined as follows:*

$$LOF_{k,\mathcal{S}}(x) = \frac{1}{|L_k(x)|} \sum_{x' \in L_k(x)} \frac{lrd_k(x')}{lrd_k(x)}.$$

Here, $\delta(x, x')$ *is the distance between two* d*-dimensional feature vectors. The LOF Predicts* -1 *for anomalous points and* $+1$ *for inlier points.*

In this Chapter, our main goal is to provide *probabilistic guarantees* on the robustness of counterfactuals to potential model changes for differential models such as neural networks. Towards achieving this goal, our objective involves: (i) introducing an abstraction that rigorously defines the class of model changes that we are interested in; and (ii) establishing a measure, denoted as $R_\Phi(x, m)$, for a counterfactual x and a given model $m(\cdot)$, that quantifies its robustness to potential model changes.

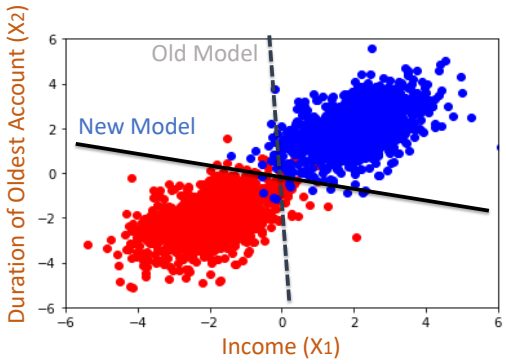


Figure 2.1: Models can often change drastically in the parameter space causing little to no change in the actual decisions on the points on the data manifold.

Here, Φ represents the hyperparameters of the robustness measure. Ideally, we desire that the measure $R_\Phi(x, m)$ should be high if the counterfactual x is less likely to be invalidated by potential model changes. We seek to provide: (i) theoretical guarantees on the validity of counterfactuals with sufficiently high value of $R_\Phi(x, m)$ with a deeper understanding of the guarantee under various assumptions; (ii) various estimators $\hat{R}_\Phi(x, m)$ and study fundamental requirements needed to ensure precise estimates; and (iii) strategies to incorporate our measure into an algorithmic framework for generating robust counterfactuals that are realistic (on the data manifold) with low cost.

2.4 Main Contributions

In this section, we first introduce our proposed abstraction of *naturally-occurring* model change and then propose a novel measure – that we call *Stability* – to quantify the robustness of counterfactuals to potential model changes. We derive a theoretical guarantee that counterfactuals that have a sufficiently high value of *Stability* will remain valid after potential *naturally-occurring* model change with high probability. But since our quantification would depend on the local Lipschitz constant around a data point, which is not known, we also examine estimators of our proposed measure.

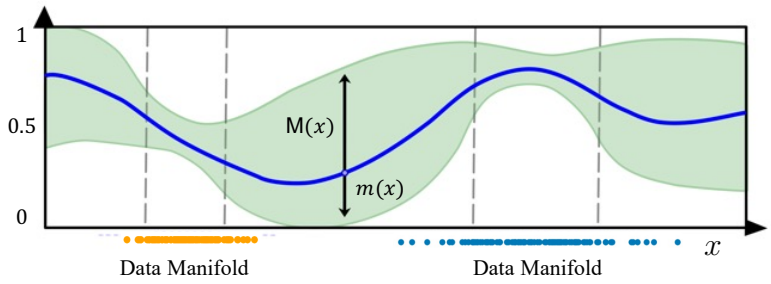


Figure 2.2: Illustrates our proposed abstraction of naturally-occurring model change: The distribution of the changed model outputs $M(x)$ (stochastic) is centered around the original model output $m(x)$. The points specifically lying on the data-manifold acting as anchors without much change as they exhibit lower variance in model outputs compared to points outside the manifold. This visualization also connects with the Rashomon effect, encapsulating the diverse yet similarly accurate models that can be learned from a given dataset.

2.4.1 Naturally-Occurring Model Change

A popular assumption in existing literature [28, 29] to quantify potential model changes is to assume that the model changes are bounded in the parameter space, i.e.,

$$\|\text{Params}(M) - \text{Params}(m)\| < \Delta \text{ for a constant } \Delta.$$

Here, Params(M) denote the parameters of the model M, e.g., weights of a neural network. However, we note that models can often change drastically in the parameter space causing little to no change in the actual decisions on the points on the data manifold (see Fig. 2.1 for an example). In this work, we avoid the bounded-model-change assumption and instead introduce the notion of a naturally-occurring model change as defined in Definition 2.5. Our abstraction allows for arbitrary model changes such that the change in predictions on points that lie on the data manifold is limited (see Fig. 2.2).

This abstraction is motivated from the observation that points residing in the data-manifold generally demonstrate reduced variance in model outputs compared to those outside the manifold. This

behavior can be attributed to the fact that during training, the model is predominantly exposed to data points from the data-manifold, leading to higher confidence in its predictions in that regions. Consequently, the model's behavior for points outside the manifold can be unpredictable (see Fig. 2.3).

Definition 2.5 (Naturally-Occurring Model Change)**.** *We make the following assumptions:*

1. $\mathbb{E}\left[M(X)|X=x\right] = \mathbb{E}\left[M(x)\right] = m(x)$ *where the expectation is over the randomness of* M *given a fixed value of* $X = x \in \mathbb{R}^d$.

2. *Whenever* $m(x)$ *is* γ_m*-Lipschitz, any updated model* $M(x)$ *is also* $\gamma-$*Lipschitz for some constant* γ*. Note that, this constant* γ *does not depend on* M *since we may define* γ *to be an upper bound on the Lipschitz constants for all possible* M *as well as* m*.*

3. $\mathrm{Var}\left[M(X)|X=x\right] = \mathrm{Var}\left[M(x)\right] = \nu_x$ *which depends on the fixed value of* $X = x \in \mathbb{R}^d$*. Furthermore,* ν_x *is small for* x *lying on the data manifold* $\mathcal{X}$*.*

Closely connected to naturally-occurring model change is the idea of the Rashomon effect, alternatively known as predictive or model multiplicity. [12, 13, 15, 39] which suggests that models can be very different from each other but have almost similar performance on the data manifold. Model multiplicity arises when models trained on the same dataset (with different weight initialization) assign varying predictions to a given sample. This is mainly because the primary objective of training is to minimize empirical risk loss. Consequently, several models can be distinctly different (even yielding opposing predictions) but still maintain comparable accuracy levels. These models are generally more confident on the data manifold, e.g., $\frac{1}{n}\sum_{i=1}^{n}|M(x_i) - m(x_i)|$ is small when the points x_i lie on the data manifold. Under the naturally-occurring model change, this holds in expectation:

Theorem 2.1 (Connection to Rashomon Effect)**.** *For points $x_1, \ldots, x_n \in \mathcal{X}$ (lying on the data-manifold) under naturally-occurring model change, the following holds:*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}|M(x_i) - m(x_i)|\right] \leq \sqrt{\nu}, \quad \textit{where} \quad \nu = \frac{1}{n}\sum_{i=1}^{n}\nu_{x_i} \tag{2.1}$$

Rashomon effect [12, 13, 15, 39] or model multiplicity typically refers to the phenomenon of a diverse models yielding similar accuracy levels on the same dataset. The Rashomon set [15] aims to characterize the entire set of models whose predictions differ by a small amount with additional constraints, e.g., models within a certain model class. We adopt a probabilistic stance on model multiplicity. Our proposed abstraction of naturally-occurring model change attempts to characterize the distribution of the models which are more likely to occur naturally rather than the entire set. Theorem 2.1 ties to the Rashomon effect by demonstrating how, under naturally-occurring model multiplicity, different models (despite their varied structures and predictions) can exhibit a surprisingly consistent performance when evaluating on points lying on the data manifold. This consistency is quantified by the expectation that the absolute difference in predictions across models is bounded, implying that a diverse set of models can indeed yield similar accuracy levels on the same dataset. Thus, Definition 2.5 is better suited over boundedness in the parameter space. Proof of Theorem 2.1 is in Appendix A.2.

Remark 2.1 (Adversarial Model Change)**.** *In contrast to naturally-occurring model change, we also introduce adversarial model change (targeted) which essentially refers to a model change that is more deliberately targeted to make a particular counterfactual invalid.*

Theorem 2.2 (Impossibility Under Adversarial Change)**.** *Given a model and counterfactual, one can always design a similar model such that the particular targeted counterfactual can be invalidated.*

The proof, provided in Appendix A.3, shows that there is a new model $M(x) = m(x)$ almost everywhere except at or around the targeted point x', i.e., $M(x') = 1 - m(x')$. Such a model could emerge from training with a poisoned data point or as a split model. These scenarios represent adversarial manipulations rather than naturally occurring model variations, and illustrate non-standard model behaviors that we distinguish from the naturally occurring model changes.

2.4.2 Measure of Robustness of a Counterfactual

Definition 2.6 (Stability)**.** *Given a model $m(\cdot)$, the stability of a counterfactual $x \in \mathbb{R}^d$ is defined as:*

$$R_{k,\sigma^2}(x, m) = \frac{1}{k} \sum_{x_i \in N_{x,k}} \left(m(x_i) - \gamma \|x - x_i\| \right), \tag{2.2}$$

where $N_{x,k}$ is a set of k points drawn from the Gaussian distribution $\mathcal{N}(x, \sigma^2 \mathrm{I}_d)$ and γ is an upper bound on the Lipschitz constant for all models $M(\cdot)$ under naturally-occurring change.

Our stability measure generalizes to any predictive class, as it can be fundamentally tied to the confidence of predicting a class (in our case class 1). For cases where a prediction needs to shift from 1 to 0, the concept can seamlessly apply by considering the logits (or softmax outputs) for predicting class 0. This could also extend to multi-class classification providing logits for each class.

Since obtaining the precise Lipschitz constant for neural networks is a complex task; hence, we operate under the assumption of a finite upper bound on the Lipschitz continuity for both our original model and changed models. This assumption might be more likely to hold if all the models belong to the same model class with roughly similar architectures. Furthermore, in practice models can also be trained using regularization to prevent their Lipschitz constant from being very high [54].

Remark 2.2 (Relaxations to local Lipschitz)**.** *While we prove our theoretical result (Theorem 2.3) with the global Lipschitz constant γ, we can relax this to local Lipschitz constants γ_x, around a given point x. This is because we sample from a Gaussian centered around the point x and hence mainly capture the variability around x. So most points will be very close to x but a few points can still lie far away. Potential extensions of our guarantees could apply to truncated Gaussian and uniform sampling methods, given their sub-Gaussian properties. This is because Lipschitz concentration inherently extends to sub-Gaussian random variables [55].*

2.4.3 Probabilistic Guarantees on Validity

To justify stability as a measure of robustness for a counterfactual to natural-occurring model changes, we provide a probabilistic guarantee on the validity of the counterfactual in Theorem 2.3.

Theorem 2.3 (Probabilistic Guarantee)**.** *Let $X_1, X_2, \ldots, X_k$ be k iid random variables with distribution $\mathcal{N}(x, \sigma^2 I_d)$ and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Suppose $|\mathbb{E}[Z|M] - \mathbb{E}[Z]| < \epsilon'$. Then, for any $\epsilon > 2\epsilon'$, a counterfactual $x \in \mathcal{X}$ under naturally-occurring model change satisfies:*

$$\Pr(M(x) \geq R_{k,\sigma^2}(x, m) - \epsilon) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma_m + \gamma)^2\sigma^2}\right).$$

Probability is over the randomness of both M and X_i's.

This stability metric (see Definition 2.6) is a way to measure the robustness of counterfactuals that are subject to natural model changes (see Definition 2.5). The first term in the metric, represented by $\frac{1}{k}\sum_{i=1}^{k} m(X_i)$, captures the average model outputs for a group of points centered around the counterfactual x. The second term, represented by $\gamma\|x - X_i\|$, is an upper bound on the potential difference in outputs of any new model on the points x and X_i (Recall the Lipschitz property of M around the

point x). Using our measure, the guarantee in Theorem 2.3 can be rewritten as:

$$\Pr\left(\frac{1}{k}\sum_{i=1}^{k} m(X_i) - M(x) \leq \frac{\gamma}{k}\sum_{i=1}^{k}\|x - X_i\| + \epsilon\right) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma+\gamma_m)^2\sigma^2}\right).$$

This form of the inequality allows for the following interpretation of Theorem 2.3: The distance between the output of the new model on an input x, i.e., $M(x)$, and the average prediction of the neighborhood of the given input by the old model, i.e., $\frac{1}{k}\sum m(X_i)$ is upper bounded by ϵ-corrected, γ multiplied average distance of the datapoints within the neighborhood of the input x, i.e., $\frac{1}{k}\sum\|x - X_i\|$.

Proof Sketch: The complete proof of Theorem 2.3 is provided in Appendix A.4. Here, we include a proof sketch. Notice that, using the Lipschitz property of $M(\cdot)$ around x, we have $M(x) \geq M(X_i) - \gamma\|x - X_i\|$ for all X_i. Thus,

$$M(x) \geq \frac{1}{k}\sum_{i=1}^{k}(M(X_i) - \gamma\|x - X_i\|) \overset{(a)}{\geq} \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - \gamma\|x - X_i\|) - \epsilon,$$

where (a) holds from Lemma 2.1 with probability at least $1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma+\gamma_m)^2\sigma^2}\right)$.

Lemma 2.1 (Deviation Bound)**.** *Let* $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$, $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$, *and* $|\mathbb{E}[Z|M] - \mathbb{E}[Z]| < \epsilon'$. *Then, under naturally-occurring change,* $\mathbb{E}[Z] = 0$. *Moreover, for any* $\epsilon > 2\epsilon'$,

$$\Pr(Z \geq \epsilon) \leq \exp\left(\frac{-k\epsilon^2}{8(\gamma+\gamma_m)^2\sigma^2}\right). \tag{2.3}$$

Proof Sketch: The proof of Lemma 2.1 leverages concentration bounds for Lipschitz functions of independent Gaussian random variables (see Lemma 2.2). Detailed proof is provided in Appendix A.4.

□

Lemma 2.2 (Gaussian Concentration Inequality)**.** *Let* $W = (W_1, W_2, \ldots, W_n)$ *consist of* n *i.i.d. random variables belonging to* $\mathcal{N}(0, \sigma^2)$*, and* $Z = f(W)$ *be a* γ*-Lipschitz function, i.e.,* $|f(W) - f(W')| \leq \gamma \|W - W'\|$*. Then:*

$$\Pr(Z - \mathbb{E}[Z] \geq \epsilon) \leq \exp\left(\frac{-\epsilon^2}{2\gamma^2\sigma^2}\right) \textit{ for all } \epsilon > 0. \tag{2.4}$$

Refer to [56] in p.125. for proof of Lemma 2.2. Our robustness guarantee (Theorem 2.3) states that $\Pr(M(x) \leq R_{k,\sigma^2}(x, m) - \epsilon) \leq \exp\left(\frac{-k\epsilon^2}{8(\gamma+\gamma_m)^2\sigma^2}\right)$ under naturally-occurring model change. For instance, a counterfactual x such that $R_{k,\sigma^2}(x, m) - \epsilon$ is greater or equal to 0.5, then $M(x)$ would also be greater than 0.5 with high probability. The term $\exp\left(\frac{-k\epsilon^2}{8(\gamma+\gamma_m)^2\sigma^2}\right)$ decays with k.

In Theorem 2.3, we assume the bound $|\mathbb{E}[Z|M] - \mathbb{E}[Z]| < \epsilon'$. In Corollary 2.1, we relax this assumption to $\Pr\left(|\mathbb{E}[Z|M] - \mathbb{E}[Z]| > \epsilon'\right) \leq \delta$, i.e., the bound is relaxed to allow a probability of δ for the deviation to exceed ϵ' (see proof in Appendix A.5). For small δ, a high stability measure implies a high probability of being valid for changed models.

Corollary 2.1. *Let* $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$ *and* $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$*. Suppose* $\Pr\left(|\mathbb{E}[Z|M] - \mathbb{E}[Z]| > \epsilon'\right) \leq \delta$*. Then, for any* $\epsilon > 2\epsilon'$*, a counterfactual* $x \in \mathcal{X}$ *under naturally-occurring model change satisfies:*

$$\Pr(M(x) \geq R_{k,\sigma^2}(x, m) - \epsilon) \geq (1 - \delta)\left(1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma_m+\gamma)^2\sigma^2}\right)\right).$$

Probability is over the randomness of both M *and* X_i'*s.*

Under certain assumptions, we are able to characterize δ. For instance, in Assumption 2.1, we build on condition (3) of Definition 2.5 by bounding the expected variance around the neighborhood of a point.

Assumption 2.1. *Let x be a point that lies on the data manifold $\mathcal{X}$. Assume that the random variable X is drawn from a Gaussian distribution $\mathcal{N}(x, \sigma^2 I_d)$. Under these conditions, we make the following assumption: $\mathbb{E}_X[\mathrm{Var}(M(X)|X)] = \mathbb{E}_X[\nu_X] \leq \alpha$, where α is a small constant. The expectations are over X, and the variance over M.*

This assumption posits that the expected variance of the changed models' prediction around the neighborhood is bounded by a small constant α. Points residing on the data-manifold generally demonstrate reduced variance in model outputs compared to those outside the manifold since the model is predominantly exposed to training data points from the data-manifold, leading to higher confidence in its predictions in those regions (see Fig. 2.3 for illustration).

Leveraging this Assumption 2.1, we introduce Lemma 2.3, which serves as the foundation for proving Theorem 2.4 which offers a comprehensive probabilistic guarantee on the validity of counterfactuals with a high Stability value on the data manifold (see Appendix A.5 for proof).

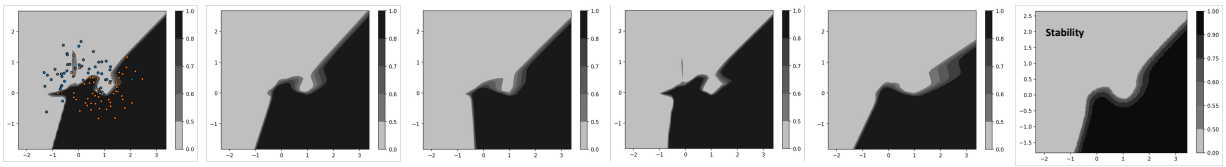
Lemma 2.3. *Let $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$ on the data-manifold and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Then, for all $\epsilon > 0$, under naturally-occurring model change and Assumption 2.1,*

$$\Pr(|\mathbb{E}\left[Z|M\right] - \mathbb{E}\left[\mathbb{E}\left[Z|M\right]\right]| \geq \epsilon) \leq \frac{\alpha}{\epsilon^2}. \tag{2.5}$$

Theorem 2.4. *Let $X_1, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$ on the data-manifold and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Then, a counterfactual $x \in \mathcal{X}$ under Assumption 2.1 and naturally-occurring model change satisfies:*

$$\Pr(M(x) \geq R_{k,\sigma^2}(x, m) - \epsilon) \geq \left(1 - \frac{\alpha}{\epsilon^2}\right)\left(1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma_m + \gamma)^2\sigma^2}\right)\right).$$

Probability is over the randomness of both M and X_i's.



(a) m (b) $M = m_1$ (c) $M = m_2$ (d) $M = m_3$ (e) $M = m_4$ (f) $\hat{R}_{k,\sigma^2}(x, m)$

Figure 2.3: Effect of stability measure on naturally-occurring model changes: (a) corresponds to the original data distribution and the trained model. (b)-(e) demonstrate some examples of changed models obtained on retraining with different weight initializations. One may notice that the model decision boundary is changing a lot in the sparse regions of the data-manifold (few data-points), possibly violating the bounded-parameter change assumption but the predictions on the dense regions of the data-manifold do not change much (in alignment with Rashomon effect). This motivates our proposed abstraction of naturally-occurring model change which allows for arbitrary changes in the parameter space with little change in the actual predictions on the dense regions of the data manifold. (f) demonstrates our proposed measure of stability $\hat{R}_{k,\sigma^2}(x, m)$ (high mean model output, low variability, *almost* like a Gaussian filter) for which we derive probabilistic guarantees on validity. In essence, we show that under the abstraction of naturally-occurring model change, the stability measure captures the reliable intersecting region of changed models with high probability. In the original model, we observe that certain non-robust regions (i.e., those caused by overfitting to certain data points in the original model) have higher local Lipschitz values and variability. Counterfactuals assigned to these regions (even if $m(x)$ is high) would be invalidated in the changed models. The stability measure, which samples around a region, penalizes these higher local Lipschitz values.

2.4.4 Estimators of Stability and their Properties

Here, we provide practical estimators of stability measure since true stability (see Definition 2.6) relies on the Lipschitz constant γ (or the local Lipschitz constant γ_x around the point x), which is often unknown. We propose two practical estimators and study their properties.

Definition 2.7 (Stability-Lipschitz Estimator)**.** *Let $N_{x,k}$ be a set of k points drawn from the Gaussian*

distribution $\mathcal{N}(x, \sigma^2 \mathrm{I}_d)$, the stability (Lipschitz Estimate) of a counterfactual $x \in \mathbb{R}^d$ is defined as:

$$\hat{R}_{k,\sigma^2}(x, m) = \frac{1}{k} \sum_{x_i \in N_{x,k}} \left(m(x_i) - \hat{\gamma}_x \|x - x_i\| \right),$$
$$\textit{where} \quad \hat{\gamma}_x = \max_{x_i \in N_{x,k}} \frac{|m(x) - m(x_i)|}{\|x - x_i\|}. \tag{2.6}$$

The Stability-Lipschitz Estimate aims to approximate the local Lipschitz constant γ_x through the term $\hat{\gamma}_x$. By design, this estimate focuses on capturing the worst-case variability in the local neighborhood of the point x.

Insight into the Stability-Lipschitz Relaxation: Through this localized assessment, our estimate offers a fine-grained, yet computationally feasible, metric for stability. However, the accuracy of this estimate is closely tied to the number of samples k drawn from the local neighborhood of x. In essence, a sufficient k is crucial for the robust approximation of the local Lipschitz constant γ_x. We formally address this requirement in Theorem 2.5, where we derive a fundamental lower bound on k to ensure that the Stability-Lipschitz estimator approximates the true stability within an ϵ error.

Theorem 2.5 (Fundamental Lower Bound on Sample Size)**.** *Let $\mathcal{M}$ be a class of all models with Lipschitz constant γ in domain $[-T, T]^d \subset \mathbb{R}^d$ and bound on the second-order partial derivatives, i.e., $\forall m \in \mathcal{M}$, $\left|\frac{\partial^2 m}{\partial x_i \partial x_j}\right| \leq \psi$ for all $x \in \mathbb{R}^d$ and $i, j \in \{1, 2, \ldots, d\}$. If,*

$$\textit{sup}_{m \in \mathcal{M}} \mathbb{E}\left[\left|\hat{R}_{k,\sigma^2}(x, m) - R_{k,\sigma^2}(x, m)\right|\right] < \epsilon,$$

then $k \geq \left(\frac{\sqrt{2\sigma^2}\, T\, \psi\, \Gamma\left(\frac{d+1}{2}\right)}{9.69 \epsilon\, \Gamma\left(\frac{d}{2}\right)} \right)^d$, where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ (Gamma function).

Theorem 2.5 highlights that the estimation of our measure is adversely affected by the curse

of dimensionality, meaning that as the dimensionality increases, so does the number of samples required for accurate estimation. This poses a computational challenge, particularly when employing gradient-based methods to identify robust counterfactuals based on stability metrics. To mitigate this computational burden, we introduce the Stability-Soft Estimator as a more efficient, albeit less accurate, alternative for estimating stability. To arrive at this estimator, we utilize the Lipschitz property, specifically, by approximating $\gamma_x||x - x_i||$ with $|m(x) - m(x_i)|$.

Remark 2.3. *A reverse statement of Theorem 2.5 would depend on the particular estimation technique. Estimating the Lipschitz constant is challenging in general, and most estimators tend to underestimate the true Lipschitz constant. This happens because even if there is a small region of the input manifold where the model has erratic behavior, the global Lipschitz constant is high and this can be missed in estimation if there are no samples collected from that small region. Proving a reverse might require additional assumptions, e.g., the Lipschitz constant is further known to be bounded or has limited variation which will be explored in future work.*

Definition 2.8 (Stability-Soft Estimator)**.** *Let $N_{x,k}$ be a set of k points drawn from the Gaussian distribution $\mathcal{N}(x, \sigma^2 \mathrm{I}_d)$, the stability variance estimator of a counterfactual $x \in \mathbb{R}^d$ is defined as follows:*

$$\hat{R}_{k,\sigma^2}(x, m) = \frac{1}{k} \sum_{x_i \in N_{x,k}} (m(x_i) - |m(x) - m(x_i)|).$$

Properties of Stability Estimators: To gain a deeper understanding of stability, we now consider some desirable properties of counterfactuals from [32], which proposed these properties for tree-based ensembles. The first property is based on the fact that the output of a model $m(x) \in [0, 1]$ is expected to be higher if the model has more confidence in that prediction.

Property 2.1. *For $x \in \mathbb{R}^d$, a higher $m(x)$ makes it less likely to be invalidated due to model changes.*

A high $m(x)$ alone does not guarantee robustness, as local variability around x can make predictions less reliable, e.g., points with high $m(x)$ near the decision boundary are more vulnerable.

Property 2.2. *An x is less likely to be invalidated if neighborhood points (x') have a high $m(x')$.*

Counterfactuals may also be more likely to be invalidated if it lies in a highly variable region. This is because the confidence of the model predictions in that region may be less reliable.

Property 2.3. *An x is less likely to be invalidated if model outputs around x have low variability.*

We recognize the insights provided by the three axiomatic properties which highlight individual aspects contributing to robustness. We note that robustness cannot be ascribed to any single property in isolation. Rather, it is that the collective integration of these properties—high confidence in predictions (*Property 1*), the reinforcement of confidence through neighborhood consensus (*Property 2*), and low variability in model outputs around a point (*Property 3*)—is essential for the robustness of a counterfactual. Our stability measures excel at collectively respecting these properties.

Given a point x, it generates a set of k points centered around x. The first term $\frac{1}{k}\sum_{x' \in N_{x,k}} m\left(x'\right)$ is expected to be high if the model output value $m(x)$ is high for x as well as several points close to x. But the mean value of $m(x')$ around a point x may not always capture the variability in that region, hence, the second term of the stability measure. In the Stability-Lipschitz estimator, the second term $\frac{1}{k}\sum_{x' \in N_{x,k}} \hat{\gamma}_x \|x - x'\|$ captures the *worst-case variability* of the model outputs in the neighborhood of x. The second term of the Stability-Soft estimator, $\frac{1}{k}\sum_{x' \in N_{x,k}} |m(x) - m(x')|$, captures the *average variability* of the model outputs around x. The variability term is only useful in conjunction with the

mean term which captures the average confidence in the neighborhood of a given point. This mean term along with the variability term make up our stability measure. Fig. 2.3 provides an example on a synthetic dataset to show the effect of our stability measure on naturally changed models realized from actual experiments by retraining with different weight initializations.

2.4.5 Generating Robust Counterfactuals using Our Proposed Measure: Stability

In this section, we examine several techniques of incorporating our proposed stability measure for generating robust counterfactuals for neural networks. We first define a counterfactual robustness test along the lines of [32].

Definition 2.9 (Counterfactual Robustness Test)**.** *A counterfactual satisfies test if:* $\hat{R}_{k,\sigma^2}(x, m) \geq \tau$.

Closest Robust Counterfactual. We focus on finding a point that satisfies the robustness test. The threshold value of τ can be adjusted based on the desired effective validity. Hence, a larger threshold would likely ensure that the new model, M, remains valid with high probability.

We propose Algorithm 1, T-Rex:I, which incorporates our measure to find robust counterfactuals on top of any preferred base method for generating counterfactuals. It evaluates the stability of the generated counterfactual and, if necessary, iteratively updates the generated counterfactual through a gradient descent process until a robust counterfactual that meets the desired criteria is obtained. We anticipate that since the robustness measure maximizes the mean value of the model prediction probabilities, it would steer toward the more favorable region. We also check for $m(x_c) \geq 0.5$ as a stopping criterion. An alternative T-Rex variant could start directly from the original instance x, aiming to find a counterfactual that is both close and robust by integrating multiple (differentiable) loss functions including a distance metric and our stability measure.

Algorithm 1 T-Rex:I: Theoretically Robust EXplanations: Iterative Version

Input: Model $m(\cdot)$, Datapoint x with $m(x) < 0.5$,
Algorithm parameters (k, σ^2, η, τ, max_steps).
Generate initial counterfactual x' using any technique.
Initialize robust counterfactual $x_c = x'$ and steps = 0.
while $\hat{R}_{k,\sigma^2}(x_c, m) < \tau$ and steps $<$ max_steps **do**
  Compute $\hat{R}_{k,\sigma^2}(x_c, m)$
  Compute gradient $\nabla_{x_c} \hat{R}_{k,\sigma^2}(x_c, m)$
  Update x_c via gradient ascent:
    $x_c = x_c + \eta \nabla_{x_c} \hat{R}_{k,\sigma^2}(x_c, m)$
  Increment steps
end while
Output x_c and exit

Remark 2.4 (Gradient of Stability)**.** *In Algorithm 1 and 3, we compute the gradient of $R_{k,\sigma^2}(x, m)$ with respect to x (not model m parameters). Such gradients w.r.t. x instead of m are also computed commonly in adversarial machine learning and also in feature-attributions for explainability. We use TensorFlow `tf.GradientTape` for automatic differentiation, which allows for the computation of gradients with respect to certain inputs.*

Robust Counterfactuals on Data Support. In certain cases, it may be desirable to generate counterfactuals from a predefined set of data points (i.e., training dataset $\mathcal{S}$. This is to remove the risk of producing unrealistic or anomalous results. Hence, we define the Robust Data Support Counterfactual.

Definition 2.10 (Robust Data Support Counterfactual)**.** *Given $x \in \mathbb{R}^d$ such that $m(x) < 0.5$, its robust nearest neighbor counterfactual $\mathcal{C}^{(\tau)}_{p,\mathcal{S}}(x, m)$ with respect to the model $m(\cdot)$ and dataset $\mathcal{S}$ is defined as another point $x' \in \mathcal{S}$ that minimizes the l_p norm $\|x - x'\|_p$ such that $m(x') \geq 0.5$ and $\hat{R}_{k,\sigma^2}(x', m) \geq \tau$.*

The closest data-supported counterfactual serves as a reliable reference, as it inherently has a high Local Outlier Factor (LOF). We propose Algorithm 2, T-Rex:NN, for finding data-supported counterfactuals. The algorithm begins by locating the K nearest neighbor counterfactuals to a given point x within the dataset $\mathcal{S}$. It then iterates through each of these candidates, evaluating them against

Algorithm 2 T-Rex:NN: Theoretically Robust EXplanations: Nearest Neighbor Version

Input: Model $m(\cdot)$, Datapoint x with $m(x)<0.5$, Dataset $\mathcal{S}$, Algorithm parameters (K, σ^2, k, τ).
Let $\mathrm{NN}_x = (x'_1, x'_2, \ldots, x'_K)$ be the K nearest neighbors to x with $m(x') \geq 0.5$,
for $x'_i \in \mathrm{NN}_x$ **do**
  Perform counterfactual robustness test on x'_i:
    Check if $\hat{R}_{k,\sigma^2}(x'_i, m) \geq \tau$
  if counterfactual robustness test is satisfied: **then**
    Output x'_i and exit
  end if
end for
Output: *No robust counterfactual found* and exit

a robustness test, $\hat{R}_{k,\sigma^2}(x', m) \geq \tau$. If a counterfactual meets this criterion, it is considered robust and the algorithm terminates.

Robust Counterfactuals on Data Manifold. Here, we focus on finding robust counterfactuals on the data manifold $\mathcal{X} \subseteq \mathbb{R}^d$ (realistic samples; see Definition 2.3). We leverage generative models to learn a lower dimensional latent representation of the data manifold in $\mathbb{R}^l$ where $l < d$. We focus on the Variational Auto-Encoders (VAEs) [57]. We designate the encoder component of the VAE, parameterized by θ, as $F_\theta : \mathbb{R}^d \to \mathbb{R}^l$ which transforms any data point $x \in \mathcal{X}$ into its corresponding latent variable $z \in \mathbb{R}^l$. The decoder denoted as $G_\phi : \mathbb{R}^l \to \mathbb{R}^d$, parameterized by ϕ, maps the latent variable back to the original data space.

Our method uses this latent space learned by a VAE for robust counterfactual search. Given that the latent space captures the data manifold, searching for counterfactuals in this representation enables us to discover instances coherent with the intrinsic data distribution and hence more plausible (higher LOF). The objective is as follows:

$$z' = \arg\min_z \ell(m(G_\phi(z)), 1) + \lambda_1 \|x - G_\phi(z)\|_p - \lambda_2 \hat{R}_{k,\sigma^2}(G_\phi(z), m).$$

Algorithm 3 T-Rex: Hybrid: Theoretically Robust EXplanations: Hybrid Version

Input: Model $m(\cdot)$, Dataset $\mathcal{S}$, Datapoint x with $m(x) < 0.5$,
Algorithm parameters (k, σ^2, η, τ, λ_1, λ_2, max_steps)
Train VAE encoder $F_\theta(\cdot)$ and decoder $G_\phi(\cdot)$ with dataset $\mathcal{S}$
Initialize $z = F_\theta(x)$
while steps $<$ max_steps **do**
  $z \leftarrow z - \eta \nabla_z (\ell(m(G_\phi(z)), 1) + \lambda_1 \|x - G_\phi(z)\|$
    $-\lambda_2 \hat{R}_{k,\sigma^2}(G_\phi(z), m))$
  steps $\leftarrow$ steps $+1$
  if $m(G_\phi(z)) > 0.5$ and $\hat{R}_{k,\sigma^2}(G_\phi(z), m) > \tau$ **then**
    Return $x' = G_\phi(z)$ and exit
  end if
end while
Return *No robust counterfactual found* and exit

Here $\ell(\cdot, \cdot)$ denotes a differentiable loss function (e.g. mean square loss, $\ell(u, v) = (u - v)^2$ or binary cross-entropy, loss $\ell(u, v) = -[v \log(u) + (1 - v) \log(1 - u)]$) that minimizes the gap between the prediction and the favorable outcome of 1, and the counterfactual returned is $G_\phi(z')$. The counterfactual lies in the data manifold since our algorithm obtains the latent encoding of our sample x using the encoder $z = F_\phi(x)$. The gradient steps are in the latent space of the encoder to minimize our overall loss function until we reach a z with robustness threshold $R_{k,\sigma^2}(G_\phi(z), m) > \tau$ on the desired side of the decision boundary. The details are in Algorithm 3: T-Rex: Hybrid.

2.5 Experiments

Here, we present experimental results to demonstrate how our proposed Algorithm 1 & 2 utilizes our stability measure to generate robust counterfactuals effectively.[1]

Datasets. We conduct experiments on several benchmark datasets, namely, HELOC [36], German Credit, Cardiotocography (CTG), Adult [37], and Taiwanese Credit [38]. These have two classes, with one class representing the most favorable outcome, and the other representing the least desirable

[1]Implementation is available at `https://github.com/FaisalHamman/TReX-Counterfactuals`

Table 2.1: Experimental results. Comparative results of counterfactual generation methods across datasets showing that our proposed T-Rex variants achieve higher validity and LOF while maintaining competitive cost.

		l_1 *based*				l_2 *based*			
	Method	COST	LOF	WI VAL.	LO VAL.	COST	LOF	WI VAL.	LO VAL.
HELOC	min Cost	**0.40**	**0.49**	38.8%	35.2%	**0.11**	**0.75**	13.5%	13.5%
	min Cost+T-Rex:I (Ours)	1.02	0.38	**98.2**%	**98.1**%	0.29	0.68	**98.5**%	**98.2**%
	min Cost+SNS	1.20	0.30	98.0%	97.8%	0.31	0.64	97.9%	97.0%
	ROAR	1.69	0.41	92.6%	91.2%	1.91	0.43	86.3 %	84.8%
	NN	**1.91**	0.80	51.1%	50.3%	**0.56**	0.80	51.1%	50.3%
	T-Rex:NN (Ours)	2.50	**0.92**	**84.0**%	**84.0**%	0.77	**0.92**	**84.0**%	**84.0**%
GERMAN	min Cost	**1.42**	**0.77**	58.8%	56.7%	**0.48**	**0.81**	26.6%	26.6%
	min Cost+T-Rex:I (Ours)	4.81	0.72	**98.0**%	96.5%	1.20	0.75	99.2%	98.7%
	min Cost+SNS	5.71	0.67	97.5%	**98.1**%	1.44	0.68	**99.9**%	**98.9**%
	ROAR	7.63	0.54	96.3%	92.3%	6.81	0.58	87.8%	85.2%
	NN	**7.05**	**1.00**	95.3%	95.4%	**2.50**	**1.00**	95.3%	95.3%
	T-Rex:NN (Ours)	10.13	**1.00**	**100**%	**100**%	3.04	**1.00**	**100**%	**100**%
CTG	min Cost	**0.21**	**0.94**	74.6%	70.2%	**0.08**	**1.00**	19.7%	14.1%
	min Cost+T-Rex:I (Ours)	1.11	0.83	**100**%	**98.8**%	0.42	0.94	**100**%	**99.7**%
	min Cost+SNS	3.34	-1.00	**100**%	98.2%	1.07	-1.00	**100**%	99.3%
	ROAR	3.68	0.64	98.7%	96.4%	1.35	0.59	98.9%	97.2%
	NN	**0.39**	**1.00**	70.5%	67.5%	**0.15**	**1.00**	70.5%	67.5%
	T-Rex:NN (Ours)	2.22	-0.33	**100**%	**100**%	1.00	-0.33	**100**%	**100**%

outcome for which we aim to generate counterfactuals. We normalize features to lie between $[0, 1]$.

Performance Metrics. Our metrics of interest are: 1.) Cost: Average l_1 or l_2 distance between counterfactuals x' and original points x. 2.) Validity (%): Percentage of counterfactuals that remain valid under the new model M. 3.) LOF: Predicts -1 for anomalous points, and $+1$ for inliers. A high average LOF essentially suggests the points lie on the data manifold and hence more realistic, i.e., *higher is better* (see Definition 2.4). We use an existing implementation to compute LOF from [58].

Methodology. We begin by training a baseline neural network model and find counterfactuals for data points with true negative predictions. To test the robustness of these counterfactual examples, we then train 50 new models (M) and evaluate the $validity$ of the counterfactuals under different model change scenarios, which include: (i) Weight Initialization (WI): Retraining new models using the same

hyperparameters but with different weight initialization by using different random seeds for each new model; and (ii) Leave Out (LO): Retraining new models by randomly removing a small portion (1%) of the training data each time (with replacement) as well as different weight initialization. This can be justified by the concept of machine unlearning, especially within the context of regulatory frameworks like the GDPR [9]. The "right to be forgotten" mandates the deletion of an individual's data upon request, which may necessitate updates to the model. These updates can affect the reliability of previously provided explanations, thereby posing a challenge to reconciling the "right to explanation" with the "right to be forgotten" [30].

Hyperparameter selection. Our findings indicate that higher k improves robustness, but comes at the cost of increased computational cost. Our choice of k=1000 also aligns with practices in the adversarial robustness literature, where similar trade-offs between performance and computational feasibility are considered. The value of σ^2 was determined by analyzing the variance of the features. In the dataset with the features between $[0, 1]$, we found that a value of σ^2=0.01 produced good results. The threshold τ is a critical aspect of our method and can be adjusted based on the desired effective validity. A higher τ value improves validity at the expense of cost. See Appendix A.7 for more details.

Baselines. We compare our approaches with established baselines. First, we find the min Cost (l_1 and l_2) counterfactual [25] and use it as our base method for generating counterfactuals. We then compare T-Rex:I to the Stable Neighbor Search (SNS) [29] and Robust Algorithmic Recourse (ROAR) [28]. We evaluate the performance of our Robust Nearest Neighbor (Algorithm 2: T-Rex:NN) against the Nearest Neighbor (NN) counterfactuals (closest data-support robust counterfactual in Definition 2.10). We choose a value of τ to get high validity and compare cost and LOF with baselines.

Results. Results for HELOC, German Credit, and CTG datasets are in Table 2.1. Observe that the min Cost counterfactual is not robust to variations in the training data or weight initialization as expected.

ROAR generates counterfactuals with high validity, albeit at the expense of a higher cost. Our proposed method, T-Rex:I, significantly improves the validity of the counterfactuals compared to the minimum cost. The T-Rex:I algorithm achieves comparable validity results to the SNS method for both types of model changes, and often accomplishes this with lower costs and higher LOF. This can be observed across all three datasets for both l_1 and l_2 cost metrics. The T-Rex:NN algorithm also significantly improves the validity of the counterfactuals compared to the traditional Nearest Neighbor (NN) method and maintains a high LOF, except for the CTG dataset with a low LOF score. This appears to be an exception rather than the norm, possibly due to the specific characteristics of the CTG dataset itself. T-Rex:NN shows competitive performance on other datasets, such as the Taiwanese Credit and Adult datasets (see Appendix A.7). It comes at a price of increased cost, but the counterfactuals are guaranteed to be realistic since they are data-supported. We observe a lower LOF score on the CTG dataset. Refer to Appendix A.7 for additional results for Adult and Taiwanese credit datasets.

Ablation. To evaluate the efficacy of our proposed stability measure, we conduct an ablation study on the German credit dataset. We first evaluate a robustness measure that solely relies on the model's prediction of the counterfactual, denoted as $r(x', m) = m(x')$. We then examine a measure that only incorporates the mean, the average predictions for k points sampled from the distribution $N(x', \sigma^2 I_d)$, denoted as $r_{k,\sigma^2}(x', m) = \frac{1}{k}\sum_{x'_i \in N_{x',k}} m(x'_i)$. We compare these with our proposed robustness measure $\hat{R}_{k,\sigma^2}(x', m)$, which also takes into account the variability around the counterfactual. The results of the ablation study, for various τ thresholds, are summarized in Table A.6 in Appendix A.7.

2.6 Discussion

We introduce an abstraction called naturally-occurring model change and propose a measure, Stability, to quantify the robustness of counterfactuals with probabilistic guarantees. We show that counterfactuals with high Stability will remain valid after potential model changes with high probability. We investigate various techniques for incorporating stability in generating robust counterfactuals and introduce the T-Rex:I, T-Rex:NN, and T-Rex:Hybrid algorithms. We also make a novel conceptual connection with the body of work on model multiplicity, further emphasizing on the models that are more likely to occur.

The naturally-occurring model changes rest on assumptions that may not apply to all models or datasets. Our stability estimators, although practically implementable, lack the same theoretical guarantees as the initial stability measure. Estimating the Lipschitz constant around a counterfactual can be computationally demanding, particularly when leveraging gradient descent to optimize stability. Though generating robust counterfactuals is a key step towards trustworthy AI, it can fall short of other important factors such as fairness [59–63]. Future research could explore links between robustness and fairness, improving the estimation of stability, or integrating Stability into training-time-based approaches for generating robust counterfactuals. Our current framework assumes a continuous space, but exploring extensions to discrete feature spaces would also be interesting.

Chapter 3: Prediction Consistency Under Model Multiplicity in Tabular LLMs

3.1 Introduction

Large language models (LLMs) are generating significant interest in high-stakes applications, e.g., finance, healthcare, particularly in few-shot classification scenarios. Since many of these sectors rely on tabular data, Tabular LLMs (TabLLMs) is emerging as a research priority [64]. Recent studies have shown that TabLLMs perform commendably in scenarios with limited training data due to their transfer learning abilities [1, 2, 65–67]. However, these models are often fine-tuned from large pre-trained models with millions of parameters on small, proprietary datasets [68, 69]. This paucity of training data and large parameter space introduces arbitrariness and inconsistency across fine-tuned model variants, raising concerns about their trustworthy adoption in high-stakes applications.

One imminent challenge for TabLLMs is *fine-tuning multiplicity*, where multiple well performing models fine-tuned from the same pre-trained LLM under slightly varying conditions (e.g., different seed, hyperparameters, or minor changes in training data) produce conflicting predictions for the same inputs. This concept is closely related to predictive multiplicity, often referred to as the Rashomon effect in the context of decision trees and neural networks [13, 15, 70]. While multiplicity has also been observed in LLMs for text classification [71], we are particularly interested in fine-tuning multiplicity in TabLLMs due to their relevance in high-stakes classification tasks. E.g., in areas like finance [72] and healthcare [73–75], arbitrary and conflicting predictions on the same input under minor model

variations can lead to confusion, reputational risk, and distrust among stakeholders.

Aside from the inherent need for predictions to be consistent to minor model variations (due to seed or hyperparameters), TabLLMs deployed by institutions may also need to be updated for various reasons, e.g., to retrain on a few additional data points [76], or even remove few data points for privacy. Regulatory frameworks like the GDPR [9] introduce the *right to be forgotten* which allows unlearning an individual's data upon request, potentially leading to model updates. These model updates could, in turn, impact previously issued predictions. Fine-tuning multiplicity also paves the way for fairwashing and explanation bias [77–79], making quantifying consistency under fine-tuning multiplicity an important and practically relevant problem.

Existing approaches to measure multiplicity in machine learning often involve retraining and ensembling multiple models [13]. However, such approaches can be computationally expensive for LLMs due to large parameter sizes. This raises a key question: *Can we **preemptively** quantify the consistency of individual predictions under fine-tuning multiplicity without actual retraining and ensembling?* To address this question, we propose a novel measure, termed *local stability*, which leverages the model's local behavior around each input data point in the embedding space to estimate the prediction's susceptibility to multiplicity. Our contributions are summarized as follows:

- **Study the intriguing nature of fine-tuning multiplicity in TabLLMs.** We first demonstrate that prediction inconsistency exists when we actually fine-tune several models from the same pre-trained model, as observed through existing multiplicity measures such as *Arbitrariness*, *Discrepancy*, *Pairwise Disagreement*, as well as two of our proposed multiplicity measures, *Prediction Variance*, and *Range* (defined in Section 3.4). We also visualize the decision boundary for several TabLLMs fine-tuned for a simple classification task and unravel an interesting "noise" pattern: unlike neural

network classifiers which typically have locally-smooth decision boundaries, Tabular LLMs show abrupt and impulsive variations (see Fig. 3.2). Thus, a model having high confidence in a prediction alone does not guarantee its consistency under fine-tuning multiplicity.

- **A measure to quantify prediction consistency under fine-tuning multiplicity.** We introduce a novel measure, termed *local stability* (see Definition 3.5), to quantify the consistency of model predictions under fine-tuning multiplicity without retraining several models. Given an input $\mathbf{x}$ and a model's prediction probability for a class c, i.e., $f_c(\mathbf{x}) \in [0, 1]$, our measure is $S_{k,\sigma}(\mathbf{x}, f_c) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_{\mathbf{x},k}} (f_c(\mathbf{x}_i) - |f_c(\mathbf{x}) - f_c(\mathbf{x}_i)|)$, where $N_{\mathbf{x},k}$ is a set of k points sampled independently from a distribution over a hypersphere of radius σ centered at $\mathbf{x}$. This measure uses the input's local neighborhood (in the embedding space) to inform the local stability, capturing both the mean model confidence in the neighborhood and the variability in confidence in that region.

- **Probabilistic guarantees on consistency over a broad class of fine-tuned models.** We provide a theoretical guarantee (see Theorem 3.1) that predictions with sufficiently high local stability will remain consistent with high probability over a *broad range of equally-well-performing fine-tuned models*. To derive this, we make some mild assumptions on the behavior of this fine-tuned model class (see Assumption 3.1). Our proof leverages Hoeffding's Inequality (see Lemma B.2).

- **Experimental results.** We show that our Local Stability measure (computed preemptively without retraining or ensembling) is quite well-aligned with the consistency of data points under actual fine-tuned multiplicity for several datasets, namely, the German Credit, Bank, Heart, Car, Diabetes, and Adult datasets [37, 80, 81]. We employ the BigScience T0 encoder-decoder model [82] and Google FLAN-T5 [83], fine-tuned via the T-Few recipe [69], and LoRA [68]. For each case, we empirically evaluate the extent of fine-tuning multiplicity, and also study how our local stability

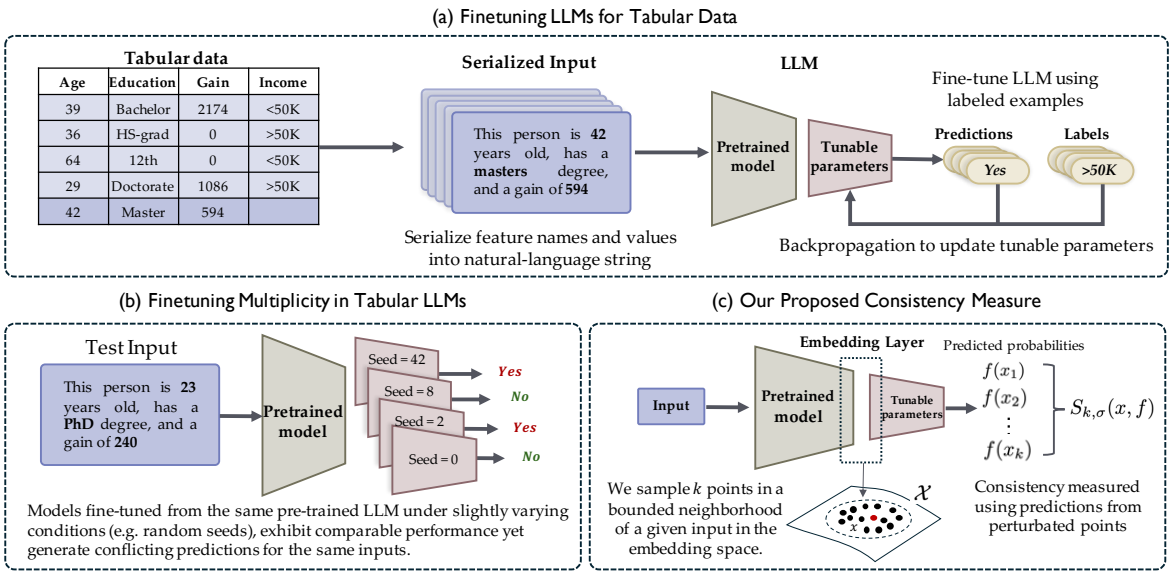


Figure 3.1: (a) illustrates the process of fine-tuning LLMs for Tabular data using few labeled examples [1, 2]. (b) demonstrates the concept of finetuning multiplicity. Models fine-tuned from the same pre-trained LLM under slightly varying conditions, such as different random seeds, can exhibit comparable performance metrics but may yield conflicting predictions for the same input. (c) introduces our proposed local stability measure designed to quantify the consistency of individual predictions without requiring the retraining of multiple models. By sampling points in a bounded neighborhood around a given input in the embedding space, the consistency measure $S_{k,\sigma}(\mathbf{x}, f)$ informs a prediction's susceptibility to multiplicity.

measure $S_{k,\sigma}(\mathbf{x}, f)$, (measured only using one model f) can preemptively capture consistency under fine-tuning multiplicity better than competing measures including prediction confidence alone.

3.2 Related Works

LLMs for tabular data is a growing area of research [65–67, 84–91]. While neural networks and gradient-boosted trees perform well with tabular data when ample labeled data is available, their effectiveness drops considerably in data-scarce scenarios. LLMs can leverage their *reasoning*, in-context learning, and pre-trained knowledge to maintain strong performance even on tiny tabular datasets [1]. [2] proposes LIFT, a method for adapting LLMs to non-language classification and regression tasks

without changing the model architecture or loss function. [1] studies the use of LLMs for zero-shot and few-shot classification of tabular data and finds that this method outperforms previous deep-learning-based approaches and is competitive with traditional baselines like gradient-boosted trees. [73] presents MediTab, a method that uses LLMs to combine different medical datasets, significantly improving predictions for patient and trial outcomes. Tabular LLMs have also been applied in other high-stakes domains [72, 74, 75, 92]. [72] presents FinPT, an LLM based approach to financial risk prediction. We refer to [93] for a detailed survey on LLMs for tabular data.

[70] introduced the idea that models can differ significantly while achieving similar average performance, known as the Rashomon effect. [13] highlighted the prevalence of arbitrary decisions in simple classification problems, calling this predictive multiplicity. [94] discuss the harms of predictive multiplicity and arbitrary decisions. Methods such as TreeFarms [95], CorelsEnum [96], and RashomonGB [97] provide tools to enumerate models in the Rashomon set for different hypothesis spaces. Efforts to leverage model multiplicity beneficially while addressing its implications have been explored by [77, 95, 98, 99]. Model multiplicity in fairness and explainability are examined by [16, 17, 29, 32, 39, 100]. [15, 101] offered a framework for measuring predictive multiplicity in machine learning models, however, this involves retraining several models, with the exception of [102] who propose a drop-out based approach to explore the Rashomon set for neural networks. Model multiplicity under fine-tuning has not been extensively studied in TabLLMs. The closest work is [71], which empirically investigates prediction arbitrariness for text classification (online content moderation). In this work, we isolate and examine a specific form of multiplicity in TabLLMs that focuses on minor model variations due to fine-tuning from the same pre-trained LLM (see Section 3.4). We leverage the embedding space of LLMs to preemptively quantify consistency under fine-tuning multiplicity without expensive retraining (see Section 3.5). There are also other alternate directions in

robustness literature that have focused on aspects other than multiplicity such as out-of-distribution generalization, adversarial examples, and uncertainty estimation [103, 104].

3.3 Preliminaries

We consider a classification task for a tabular dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where each $\mathbf{x}_i$ is a d-dimensional feature vector (rows of a tabular input), and each label y_i is binary, $y_i \in \{0, 1\}$. We focus on n-shot classification where a pre-trained model is fine-tuned on a limited number of n examples.

Serialization of Tabular Data for LLMs. To apply LLMs to tabular data, it is crucial to transform the data into a natural text format. This process, known as serialization, involves converting the table rows into a text string that includes both the column names and their corresponding values [1, 2, 65, 105]. The resultant serialized string is combined with a task-specific prompt to form the input for the LLM. There have been various proposed methods for serialization, and this is still a topic of active research [105]. Among the serializations we have examined are: list template (a list of column names and feature values), and text template (`The <column name> is <value>`). The training process uses the natural-language outputs of the LLM, mapped to valid classes in the target space, as part of fine-tuning (see Fig. 3.1). To clarify, table values are serialized into serialize($\mathbf{x}$) and then transformed into a format understandable by the LLM, tokenize(serialize($\mathbf{x}$)), which is an embedding. Since these transformations are one-to-one mappings, we denote the embedded form of input $\mathbf{x}$ also as $\mathbf{x} \in \mathcal{X}$ to represent $\mathbf{x}$ in the embedding space. This allows us to simplify the notation and directly use $\mathbf{x}$ to refer to the input values in the embedding space.

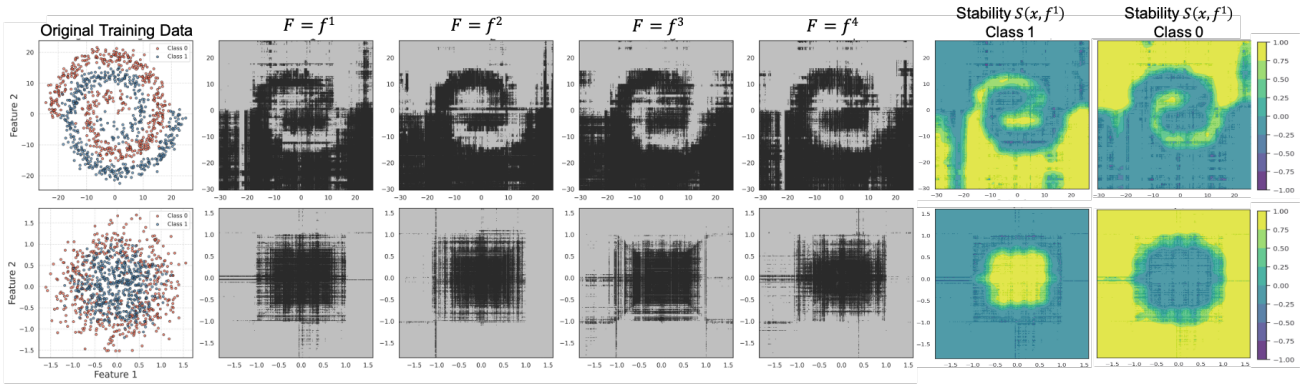


Figure 3.2: Decision boundaries for multiple fine-tuned models of an LLM on synthetic datasets. We fine-tuned several models by only changing the random training seed. All models achieve comparable training loss and accuracy, yet they converge to different functions, exhibiting intriguing noisy patterns (a phenomenon absent in models like neural networks which are typically locally-smooth). Interestingly, these noisy behaviors appear even in regions where the model is expected to confidently predict a specifc class. Observe the location and shape of these noisy patterns vary unpredictably across the various fine-tuned models, making them a possible factor contributing to prediction multiplicity. This highlights that model predictions alone may be unreliable and motivates our perturbation-based approach to quantify multiplicity. The last two plots illustrate the local stability measure applied to model f^1 across classes 0 and 1, i.e., $S(\cdot, f^1)$. The local stability measure effectively highlights regions where predictions are reliable (indicated by bright yellow color) and areas where predictions may be unstable.

3.4 Multiplicity in Fine-Tuned Tabular LLMs

Let $f(\cdot) = [f_1(\cdot), f_2(\cdot), \ldots, f_C(\cdot)] : \mathcal{X} \to \Delta^C$ denote an LLM that performs multi-class classification over C classes, where Δ^C is the C-dimensional probability simplex (e.g., softmax outputs). Let $\mathcal{F}$ denote a broad class of equally-well-performing fine-tuned models (a set of competing fine-tuned models as measured by the accuracy), i.e, $\mathcal{F}_\delta = \{f : err(f) \leq err(f_0) + \delta\}$ where $err(f_0) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\hat{f}_0(\mathbf{x}_i) \neq y_i]$ for a reference model f_0 (with satisfactory accuracy) and test dataset with N examples. Here, $\hat{f}(\mathbf{x}) = \arg\max_{c \in [C]} f_c(\mathbf{x})$ denotes the predicted label. This is a set of fine-tuned models that perform just as well as the reference baseline classifier f_0, where $\delta \in (0, 1)$ is the error tolerance. The appropriate choice of δ is application-dependent [13].

Fine-tuning Multiplicity. We study the nature of multiplicity that arises in LLMs when fine-tuned for tabular tasks. To illustrate fine-tuning multiplicity, we conduct experiments using synthetic 2D data (see Fig. 3.2). We fine-tune several competing models using the text template and varying only the random training seed. We reveal that fine-tuning LLMs on such non-language tasks exhibit noisy and non-smooth decision boundaries, even in regions where the model is expected to confidently predict a specific class. We hypothesize that this noisy behavior is likely because LLMs are optimized for capturing complex language structures. When fine-tuned on tabular data tasks, which often involve both text and numeric values, LLMs leverage their pre-trained knowledge but still exhibits instabilities.

Evaluating Fine-tuning Multiplicity. To evaluate the extent of multiplicity, we introduce specific empirical metrics that assess how predictions may vary across different fine-tuned models.

Definition 3.1 (Arbitrariness [71])**.** *Arbitrariness over set $\mathcal{F}_\delta$ measures the extent of conflicting predictions across the model space for a given set of inputs $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. It is defined as:*

$$A_\delta = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\exists f, f' \in \mathcal{F}_\delta, : \hat{f}(\mathbf{x}_i) \neq \hat{f}'(\mathbf{x}_i)]. \tag{3.1}$$

Arbitrariness generalizes the *Ambiguity* measure which computes the fraction of points where at least one model in $\mathcal{F}_\delta$ disagrees with a reference model [13]. Abitrariness measures the percentage of points that receive conflicting predictions from any two models within the set $\mathcal{F}_\delta$. Arbitrariness can also be defined on a single input, i.e., $A(\mathbf{x}_i) = \mathbb{I}[\exists f, f' \in \mathcal{F}_\delta, : \hat{f}(\mathbf{x}_i) \neq \hat{f}'(\mathbf{x}_i)]$.

Definition 3.2 (Discrepancy)**.** *Discrepancy quantifies the maximum proportion of conflicting predic-*

tions between a reference model and any competing model in the set:

$$D_\delta(f_0) := \max_{f \in \mathcal{F}_\delta} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[\hat{f}(\mathbf{x}_i) \neq \hat{f}_0(\mathbf{x}_i)]\right). \tag{3.2}$$

Discrepancy measures the maximum number of predictions that could change if a reference model is replaced with a competing model. This means that, in practice, altering multiple predictions requires that all conflicting predictions come from a single competing model.

Definition 3.3 (Pairwise Disagreement [77])**.** *Pairwise Disagreement is the proportion of instances where pairs of model predictions within the competing set disagree:*

$$PD_\delta(\mathbf{x}) := \frac{1}{|\mathcal{F}_\delta|(|\mathcal{F}_\delta| - 1)} \sum_{f^i, f^j \in \mathcal{F}_\delta, f^i \neq f^j} \mathbb{I}[\hat{f}^i(\mathbf{x}) \neq \hat{f}^j(\mathbf{x})]. \tag{3.3}$$

Since existing measures of multiplicity focus on predicted labels, we propose two more nuanced measures that leverage the predicted probabilities of model outputs:

Definition 3.4 (Prediction Variance)**.** *PV measures the variability of the model outputs for a given input $\mathbf{x}$ and class c across different models in the set $\mathcal{F}_\delta$:*

$$PV_\delta(\mathbf{x}) := \frac{1}{|\mathcal{F}_\delta|} \sum_{f \in \mathcal{F}_\delta} (f_c(\mathbf{x}) - \frac{1}{|\mathcal{F}_\delta|} \sum_{f' \in \mathcal{F}_\delta} f'_c(\mathbf{x}))^2. \tag{3.4}$$

Unlike threshold-based measures, Prediction Variance captures variability in predicted probabilities. We also define *Prediction Range* to quantify the maximum spread in predicted probabilities: $\text{PR}_\delta(\mathbf{x}) := \max_{f \in \mathcal{F}_\delta} f_c(\mathbf{x}) - \min_{f \in \mathcal{F}_\delta} f_c(\mathbf{x})$ (also see [101]). For brevity, from now on we use $f(\mathbf{x})$ to refer to the predicted probability for the specific class of interest (typically the predicted class for $\mathbf{x}$),

rather than the full vector of probabilities. Thus, $f(\mathbf{x}) \in [0, 1]$ will be a scalar.

3.5 A Novel Measure to Preemptively Capture Prediction Consistency

Our objective is to define a measure, denoted as $S(\mathbf{x}, f)$, for an input $\mathbf{x}$ and a given fine-tuned model f, that would preemptively quantify the consistency of the prediction $\hat{f}(\mathbf{x})$ over a broad class of equally-well-performing fine-tuned models. We desire that the measure $S(\mathbf{x}, f)$ should be high if the predictions for the input $\mathbf{x}$ is consistent across this broad class of fine-tuned models (see Fig. 3.1).

Candidate Measure: Prediction confidence $S(\mathbf{x}, f) := f(\mathbf{x})$. While the prediction probability of a model $f(\cdot)$ offers insights into its confidence in predicting a given class, they are insufficient for assessing consistency under fine-tuning multiplicity (see Table 3.2, Fig. 3.3, i.e., data point with high $f(\mathbf{x})$ or confidence can still be susceptible to multiplicity). In our synthetic data experiments (see Fig. 3.2), we also observe that noisy behaviors emerge in regions where the model should be confident in its predictions, leading to conflicting outcomes across various fine-tuned models. This indicates that relying solely on an input $\mathbf{x}$ may not provide a reliable assessment of consistency. To address this, we propose a perturbation-based approach that leverages the local neighborhood around the input $\mathbf{x}$ in the embedding space, ultimately leading to our measure of local stability.

3.5.1 Proposed Measure: Local Stability

Definition 3.5 (Local Stability)**.** *For a given data point $\mathbf{x}$, let $f(\mathbf{x})$ represent the predicted probability (e.g., softmax logits) from a model f. The local stability is defined as:*

$$S_{k,\sigma}(\mathbf{x}, f) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_{\mathbf{x},k}} f(\mathbf{x}_i) - \frac{1}{k} \sum_{\mathbf{x}_i \in N_{\mathbf{x},k}} |f(\mathbf{x}) - f(\mathbf{x}_i)|,$$

$N_{\mathbf{x},k}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_k\}\subset B(\mathbf{x},\sigma)=\{\mathbf{x}'\in\mathcal{X}|\|\mathbf{x}'-\mathbf{x}\|_2<\sigma\}$ *is a set of* k *points sampled independently from a distribution over a hypersphere of radius* σ *centered at* $\mathbf{x}$.

Our local stability measure is tied to the confidence in predicting a specific class (i.e., the probability values derived from softmax logits), and not on the predicted labels. It quantifies the stability of a model's predictions using the local neighborhood around an input $\mathbf{x}$. The first term, $\frac{1}{k}\sum_{\mathbf{x}_i} f(\mathbf{x}_i)$, captures the average confidence of the model in this region. The second term, $\frac{1}{k}\sum_{\mathbf{x}_i}|f(\mathbf{x})-f(\mathbf{x}_i)|$, penalizes variability by measuring how much the predictions fluctuate within the neighborhood. By subtracting this variability from the local average confidence, the measure ensures that high scores are assigned to predictions with high local neighborhood prediction confidence and low variability. This formulation is motivated by our observations on synthetic data, where models exhibited irregular, non-smooth decision boundaries despite high confidence in certain regions (see Fig. 3.2). See App. B.1 for more intuitions and properties of stability measure.

3.5.2 Theoretical Guarantees on Consistency

We present theoretical insights that motivate our proposed stability measure $S_{k,\sigma}(\mathbf{x},f)$, in quantifying the consistency of predictions across a broad class of fine-tuned models.

Let $\bar{f}(\cdot;\bar{W})$ represent a target function and $f_0(\cdot;W)$ denote a pre-trained model. Parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) adds a low-rank updates into the weight matrices of a frozen pre-trained model to effectively approximate the target function [68]. In this framework, the fine-tuned model is represented as $F(\cdot;W+\Delta W)$, where the low-rank weight updates ΔW are added to the frozen model. We let the fine-tuned model be a random variable, $F(\cdot;W+\Delta W)$, where the randomness arises from the distribution of low-rank weights $\Delta W\in\mathcal{W}$.

For clarity, we use capital letters (e.g., F, X_i, Z) to denote random variables, while lowercase letters (e.g., $\mathbf{x}_i, f, \epsilon$) indicate specific realizations. For brevity, we omit the weight parametrization.

Assumption 3.1. *Let $\bar{f}$ and F denote the target and adapted models, respectively. We assume:*

• $\mathbb{E}[F(X)|X = \mathbf{x}] = \mathbb{E}[F(\mathbf{x})] = \bar{f}(\mathbf{x})$*, i.e., $F(\mathbf{x})$ is an unbiased estimator of $\bar{f}(\mathbf{x})$.*

• $\mathbb{E}_X[\|F(X) - \bar{f}(X)\| \mid F = f] \leq \alpha$*, and the expected norm of its gradient $\mathbb{E}_X[\|\nabla(F - \bar{f})(X)\| \mid F = f] \leq t$, where X is drawn independently from a distribution over the hypersphere $B(\mathbf{x}, \sigma)$.*

• *F and $\bar{f}$ are twice differentiable with Hessians bounded by L, i.e., $\|\nabla^2 F(\mathbf{x})\|, \|\nabla^2 \bar{f}(\mathbf{x})\| \leq L$.*

Our Assumptions are motivated by recent work [106] which provides key theoretical insights into the expressive power of LoRA adaptation for transformer models. Specifically, we assume that the fine-tuned model $F(\mathbf{x})$ is an unbiased estimator of the target function $\bar{f}(\mathbf{x})$, meaning that across fine-tuned variations, the expectation of the model remains centered around the true function. Their results also establish the expected error over a bounded region, $\mathbb{E}_X[\|F(X) - \bar{f}(X)\||F = f] \leq \alpha$. Under certain conditions, they show the existence of adapter weights ΔW such that $\alpha \to 0$. Additionally, we assume that the expected gradient norm is bounded by t, implying that while gradients may vary across fine-tuned models, they remain close to the target function in expectation over a local region.

Theorem 3.1 (Probabilistic Guarantee)**.** *Given $\mathbf{x}$, a target model $\bar{f}$ and stability measure $S_{k,\sigma}(\mathbf{x}, F)$. Under Assumption 3.1, and for all $\epsilon > \epsilon'$, where $\epsilon' = 2(\alpha + t\sigma) + \mathcal{O}(L\sigma^2)$. We have:*

$$\Pr\left(\bar{f}(\mathbf{x}) \geq S_{k,\sigma}(\mathbf{x}, F) - \epsilon\right) \geq 1 - \exp\left(\frac{-k\epsilon^2}{32}\right) \tag{3.5}$$

Theoretical Guarantee Interpretation. Our stability measure $S(\mathbf{x}, f)$ provides a probabilistic guarantee that if a data point $\mathbf{x}$ has a sufficiently high local stability score on a random model F, sampled

from a broad class of equally-well-performing fine-tuned models, then the prediction on the target model $\bar{f}(\mathbf{x})$ will be at least $S(\mathbf{x}, f) - \epsilon$ with high probability. For example, if $S(\mathbf{x}, f) = 0.8$, we can be confident that $\bar{f}(\mathbf{x})$ will be at least $0.8 - \epsilon$ with *high* probability (i.e, the prediction will remain on the positive predicted side). This implies that high local stability scores are indicative of consistent predictions. The probability of the bound holding increases exponentially with the sample size k. Conversely, a low stability score does not provide significant information about the prediction's behavior, as it does not guarantee a lower bound on the prediction.

Goodness of Model Class. The term ϵ' is indicative of the quality or goodness of the fine-tuned model class. A small ϵ' indicates a well-behaved model class, suggesting that different fine-tuned models produce similar outputs in expectation within the local neighborhood of $\mathbf{x}$ even if predictions might vary for a given data point. Similar behavior is visualized in Figure 3.2, where, despite the presence of noisy variations in the decision boundaries, the local predictions around a given point remain relatively consistent across models. This behavior is expected since these models are derived from the same pre-trained model and trained to achieve similar accuracy on the dataset. In this case, *our local stability measure provides an informative lower bound on the predictions $\bar{f}(\mathbf{x})$ with a certifiably small gap*.

Conversely, a large ϵ' indicates a more erratic model class. In this case, our bound becomes less informative, and the local stability measure might perform poorly for a given point. We interpret our results as follows: The model class is not well-behaved; thus, one cannot certify a small gap between $\bar{f}(\mathbf{x})$ and our proposed measure. We do not provide guarantees for all types of model changes, as this would be challenging with only a single model. For example, if fine-tuned models do not achieve sufficient accuracy, encounter significant variations in hyperparameter choices, or large changes in the training data, ϵ' is likely to be large. Our focus is on multiplicity that arises due to randomness in training, such as changes in the training seed or minor adjustments in training settings (i.e., a broad

class of equally-well-performing fine-tuned models). In our evaluations, we do not assume any specific values for ϵ' and consider regular fine-tuned models without imposing any theoretical constraint. The proof of Theorem 3.1 is provided in App. B.2.

3.6 Experiments

In this section, we experiment across different datasets to *(i)* quantify the prevalence of fine-tuning multiplicity in TabLLMs, and *(ii)* validate the effectiveness of our proposed measure in quantifying the consistency of predictions over a broad range of equally-well-performing fine-tuned models.

Datasets and Serialization. We experiment on the Diabetes [80], German Credit [81], Bank [107], Heart, Car, and Adult datasets [37], serialized using the Text Template, i.e., tabular entry is converted into a natural language: `The <column name> is <value>`. This approach helps align the inputs with the training distribution of LLMs, enhancing their performance in few-shot scenarios [1, 2].

Models and Fine-tuning Methods. We use the BigScience T0 [82] and Google FLAN-T5 [83] encoder-decoder models as our pretrained LLMs. T0 is specifically pre-trained for zero-shot generalization through multitask learning. FLAN-T5 is instruction fine-tuned on a diverse range of tasks, achieving strong performance in few-shot settings. These make both models well-suited for our experiments. For fine-tuning, we adopt the T-Few recipe [69], known for its effectiveness in few-shot learning, and LoRA [68]. Detailed setup can be found in App. B.3.

Evaluating Extent of Fine-tuning Multiplicity. We measure the extent of fine-tuning multiplicity across the various datasets and fine-tuning methods, we use the multiplicity evaluation metrics (see Section 3.4). To evaluate these multiplicity metrics across our datasets, we fine-tune 40 models on Tfew recipe and LoRA using different random seeds and test on a sample set.

Table 3.1: Evaluated Multiplicity for Different Datasets and Number of Shots on BigScience T0. Evaluated on 40 fine-tuned models on T-Few recipe using different random seeds. Multiplicity observed in predictions across different fine-tuned model, even when models exhibit similar accuracy (in this setting $\delta = 0.02$). Fine-tuning using LoRA achieves results in the same ballpark (see LoRA Table B.1 in App. B.3)

Dataset	No. Shots	Multiplicity Evaluation Metrics (BigScience T0)					
		Arbit. Disc.		Avg. Pair. Disag.	Avg. Pred. Variance	Avg. Pred. Range	Avg. Model Accuracy
Adult	64	10%	9%	7%	0.01	0.10	83%
	128	10%	7%	8%	0.01	0.10	84%
	512	11%	8%	7%	0.01	0.12	85%
German	64	18%	10%	6%	0.01	0.20	71%
	128	17%	11%	6%	0.01	0.16	71%
	512	23%	12%	7%	0.02	0.23	72%
Diabetes	64	29%	18%	10%	0.04	0.31	71%
	128	13%	17%	11%	0.03	0.13	72%
	512	16%	16%	10%	0.02	0.18	78%
Bank	64	11%	9%	6%	0.01	0.31	66%
	128	15%	8%	7%	0.03	0.22	75%
	512	14%	8%	7%	0.02	0.16	81%
Heart	64	6%	4%	2%	0.01	0.05	78%
	128	9%	4%	3%	0.01	0.10	83%
	512	18%	7%	5%	0.01	0.19	82%
Car	64	19%	10%	6%	0.01	0.18	81%
	128	16%	7%	5%	0.01	0.14	86%
	512	8%	4%	2%	0.01	0.09	94%

• We evaluate multiplicity on the *BigScience T0* model fine-tuned using *T-Few* (see Table 3.1).

• We evaluate multiplicity on *BigScience T0* fine-tuned using *LoRA* (see Table B.1 in App. B.3).

• We evaluate multiplicity on *FLAN-T5* model fine-tuned using *T-Few* (see Table B.2 in App. B.3).

Comparing Local Stability Measure to Evaluated Multiplicity. We assess the utility of our proposed stability measure $S_{k,\sigma}(\mathbf{x}, f)$ in informing the presence of fine-tuning multiplicity. This utility is measured using the Spearman correlation coefficient (see Definition B.1), between our stability $S_{k,\sigma}(\mathbf{x}, f)$ (estimated on just one model) and the evaluated multiplicity (evaluated on several finetuned models), e.g., Spearman$(S_{k,\sigma}(\mathbf{x}, f), PV_\delta(\mathbf{x}))$ across the test set. Here, our local stability measure is taken with respect to the model's predicted class for $\hat{f}(\mathbf{x})$.

Baselines. For comparison, we include the following baselines: *1) Prediction probability* $f(\mathbf{x})$ which

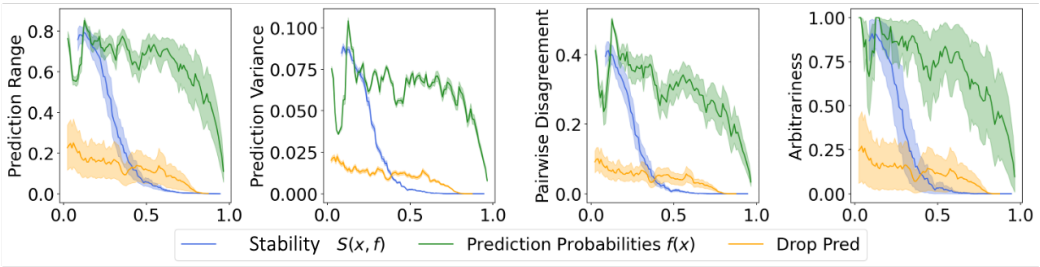


Figure 3.3: Evaluated multiplicity (assessed on 40 retrained models) versus our stability measure, predicted probabilities, and drop-out method (evaluated on one model) for the 128-shot setting on the Adult dataset. The plots demonstrate that high local stability values correspond to low multiplicity across various multiplicity evaluation metrics. Also, observe that high predicted probability values (i.e., high prediction confidence) do not imply low multiplicity. Our stability measure provides better insight into the multiplicity of predictions compared to the predicted probabilities or drop-out prediction. App. B.3 for visualizations on other datasets.

measures the confidence of the model in predicting a given class. *2) Binary Drop-Out Method* [102]: Since there are no other baselines, we adapt this Drop-out method for TabLLMs. This method drops random weights of the model to explore models in the Rashomon set (i.e., set of competing models) without retraining several models. For a fair comparison, we compare our method (sampling k points in the neighborhood of our data point in the embedding space, and computing the local stability measure) to theirs (averaging the predictions of k models with different dropped-out weights). Note that these require the same number of inferences, hence complexity for both methods are around the same. Here are the experiments we conducted:

• We plot the evaluated multiplicity against our stability measure, predicted probabilities, and the drop-out method. See Figure 3.3 for illustration on the Adult 128 shot (BigScience T0 model). For other dataset refer to Figure B.2, B.3, B.4 in App. B.3.

• We compute the absolute spearman correlation between the stability measures and various multiplicity evaluation metrics (128-shot setting on all datasets presented in Table 3.2). Results on BigScience T0 model with 64 and 512 shots are presented in Table B.3 in App. B.3. Results for FLAN-T5 model are presented in Table B.4 in App. B.3.

Table 3.2: Absolute Spearman Correlation between the stability measure and various multiplicity evaluation metrics for 128 shots on the datasets. In most cases, our stability measure $S_{k,\sigma}(x, f)$ shows a higher correlation with these multiplicity measures compared to predicted probabilities and drop-out method, indicating that the stability measure $S_{k,\sigma}(x, f)$ better informs about the multiplicity than other measures. See full Table B.3 with 64 and 512 shot cases in App. B.3.

Dataset	Number of Shots	Measure	Arbit.	Pairwise Disag.	Prediction Variance	Prediction Range
Adult	128	Pred. Prob.	0.67	0.62	0.30	0.54
		Drop-Out	0.74	0.83	0.69	0.81
		Stability	**0.80**	**0.96**	**0.84**	**0.91**
German	128	Pred. Prob.	**0.57**	**0.57**	0.86	0.86
		Drop-Out	0.50	0.56	0.74	0.84
		Stability	0.54	0.54	**0.87**	**0.87**
Diabetes	128	Pred. Prob.	0.88	0.93	**0.93**	**0.95**
		Drop-Out	0.89	0.92	0.92	0.94
		Stability	**0.92**	**0.95**	**0.93**	**0.95**
Bank	128	Pred. Prob.	0.54	0.57	0.73	0.62
		Drop-Out	0.62	0.70	0.75	0.51
		Stability	**0.79**	**0.84**	**0.87**	**0.86**
Heart	128	Pred. Prob.	0.61	0.46	0.50	0.26
		Drop-Out	0.64	0.76	0.74	0.83
		Stability	**0.89**	**0.90**	**0.97**	**0.87**
Car	128	Pred. Prob.	0.56	0.26	0.29	0.01
		Drop-Out	0.63	0.66	0.57	0.52
		Stability	**0.97**	**0.91**	**0.93**	**0.94**

Ablations and Hyperparameter Selection. Theorem 3.1 indicate that increasing the sample size k exponentially improves the probability that the stability guarantee holds. However, this also increases the computational cost of model inference. We use $k = 30$, the maximum number that fits into one inference pass on the GPU.

For the *neighborhood radius* σ, we sampled perturbed points from a truncated Gaussian distribution with a variance of 0.01, which consistently performed well across all experiments. To guide the choice of σ, we suggest the following data-driven approach. (1) Compute Pairwise Distances: For all training samples, calculate the median distance d_{med} between each point and its k-nearest neighbors

(e.g. $k = 5$) in the embedding space. (2) Set σ as a fraction of d_{med} (e.g., $\sigma = 0.1 d_{med}$). This captures the natural scale of the data while ensuring perturbations stay within the local neighborhood. This mirrors neighborhood-scale hyperparameters used in clustering, kernel methods [108, 109], and certified robustness [110, 111], which similarly rely on training data pairwise distances to set their parameters.

We used error tolerance $\delta = 0.02$ (2% margin of accuracy deviation). Evaluating multiplicity by refining multiple models is computationally expensive. Thus, we limited our study to 40 models. For the drop-out rate in the baseline, we use $p = 0.1$ following the recommendation in [102]. To evaluate the impact of varying key parameters, we conducted the following ablation studies:

• We perform an ablation study on the sample size k, observing improved performance with increasing k. Detailed results are provided in Table B.6 in App. B.3.

• We explore the effect of varying the neighborhood radius σ. Results of this ablation study are summarized in Fig. B.5 and Table B.7 in App. B.3. Best performance is observed at $\sigma = 10^{-2}$. When σ is too small (10^{-4}), we sample (almost) the same points and our local stability measure is not more informative than the prediction probability. When σ is too large (10^{-1}), information about the data point is lost.

• We also evaluate the Drop-Out method with varying drop-out rates $p \in \{0.01, 0.1, 0.2, 0.5\}$. The correlation values between evaluated multiplicity and the stability measures for the 512-shot setting on the Diabetes dataset are summarized in Table B.8 in App. B.3. Our stability measure outperforms the dropout method for all p values.

• To assess the contribution of the variability term in our stability measure, we compare it to two baselines that capture only local variability: (i) absolute deviation $S_1(\mathbf{x}) = \frac{1}{k}\sum |f(\mathbf{x}_i) - f(\mathbf{x})|$, and (ii) squared deviation $S_2(\mathbf{x}) = \frac{1}{k}\sum (f(\mathbf{x}_i) - f(\mathbf{x}))^2$. As shown in Table B.5 in App. B.3, both alternatives yield consistently weaker correlations with multiplicity metrics, highlighting the importance of

Table 3.3: Correlations and runtimes on the Adult dataset (128-shot) (100 finetuned models with an overall training time of 456 mins). Train time refers to the total time required to train the models needed; Evaluation time includes inference and computation time of the method over the entire test set. Stability achieves high correlations with multiplicity metrics at lower computational cost.

Measure	**Arbit.**	**Pairwise Disag.**	**Pred. Var.**	**Pred. Range**	**Train Time**	**Eval. Time**
Re-training	1.00	1.00	1.00	1.00	456 mins	94.7 mins
Pred. Prob.	0.63	0.61	0.39	0.63	4.56 mins	0.51 mins
Drop-Out	0.79	0.78	0.70	0.86	4.56 mins	102 mins
AWP	0.65	0.71	0.55	0.72	4.56 mins	977.6 mins
Stability	0.81	0.96	0.80	0.93	4.56 mins	19.4 mins

incorporating both local mean and variability terms.

• To evaluate the applicability of our stability measure beyond LoRA-based fine-tuning, we conduct an ablation using two alternative tuning strategies: Prompt Tuning [112] and Prefix Tuning [113]. As shown in Table B.9 in App. B.3, our Stability measure continues to correlate with multiplicity, though the correlations are somewhat weaker than in the LoRA case, likely due to the limitations of these tuning methods, which are known to be less effective than LoRA in few-shot settings.

Computational Efficiency. We compare the computational requirements of our Stability measure against, retraining, dropout-based, prediction probability, and the Adversarial Weight Perturbation (AWP) [15] in terms of both training and evaluation runtimes. Table 3.3 summarizes the cumulative training time and the total evaluation time over the Adult test set. Figure 3.4 plots each method's total runtime and correlation with multiplicity metrics. We found AWP to be expensive for LLMs since each gradient-optimization step requires full forward passes on the test set to enforce Rashomon-set constraints, incurring heavy inference and gradient-computation costs. Dropout requires a prior check to ensure all dropped-out models (the models to be aggregated) are in the competing model set, hence our method would be more computationally efficient under the same k. Stability achieves the best runtime–correlation trade-off compared to baselines (see Fig. 3.4).

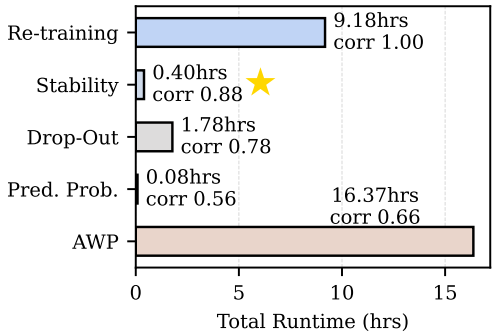


Figure 3.4: Total runtime across the Adult test dataset. Our proposed method (Stability) achieves significantly lower runtime compared to the re-training and baselines while maintaining strong average correlation with multiplicity evaluation metrics.

Table 3.4: Mean and standard deviation of stability values for correctly vs. incorrectly classified data points on the Hospital dataset. Stability achieves a larger separation between correct (0.8710) and incorrect (0.5729) data points than baselines, suggesting it is better at discriminating against unreliable predictions.

Method	**Correct**		**Incorrect**	
	Mean	Std	Mean	Std
Stability	0.8710	0.1465	0.5729	0.1458
Pred. Prob.	0.8994	0.2160	0.7965	0.2256
Drop-Out	0.8190	0.2832	0.7217	0.1929

Scalability to Large Datasets. We evaluate our method on the Hospital Readmission dataset [114] ($\sim$100k data points, 50+ features). Retraining 40 models to evaluate multiplicity would takes over 5 days (3.5 hrs/model), while our method requires only a single model and a fast forward-pass sampling step. We train one model and compare Stability and baselines across correctly and incorrectly classified test points (see Table 3.4). Stability achieves a larger separation between correct and incorrect data points than baselines. We can also use our measure to analyze data points that are both confident and stable, or identify those that appear confident but are actually unstable. In Table 3.5, we grouped predictions based on their confidence and stability values. Observe that while 41% of the predictions were both confident and stable, a notable 20% of the predictions were confident but unstable, indicating that high confidence alone is not enough.

Table 3.5: Breakdown of test predictions by confidence and stability (threshold = 0.75). 41% of predictions are both confident and stable, while a significant 20% are confident yet unstable—revealing cases where high confidence masks unreliability and underscoring the value of our stability measure.

Pred. Prob.	**Stability**	**% Test**	**Description**
High (≥ 0.75)	High (≥ 0.75)	41%	Confident & Stable
High (≥ 0.75)	Low (< 0.75)	20%	Confident but Unstable
Low (< 0.75)	High (≥ 0.75)	22%	Unconfident but Stable
Low (< 0.75)	Low (< 0.75)	17%	Unconfident & Unstable

3.7 Discussion

Our multiplicity evaluation metrics, summarized in Table 3.1, B.1, and B.2 reveal significant variability in model predictions across different fine-tuned variants, even when they exhibit similar accuracy. This multiplicity is not captured by merely examining predicted probabilities, as predictions with high confidence can still be susceptible to multiplicity (see Fig. 3.3). Our stability measure, $S_{k,\sigma}(\mathbf{x}, f)$, was compared with the prediction probabilities $f(\mathbf{x})$. The results, presented in Table 3.2, B.3, and B.4 demonstrate that our stability measure consistently shows mainly higher correlation with multiplicity metrics across all models and datasets compared to prediction probabilities and drop-out method. This indicates that $S_{k,\sigma}(\mathbf{x}, f)$ is more informative than the baselines in informing the fine-tuning multiplicity. The drop-out method is however better than the prediction probabilities alone. We hypothesize that our method is more suitable for LLMs because the embedding space of LLMs is significantly smaller than the parameter space (possibly more informative also). The drop-out method might need significantly more inferences to compete due to this.

We study the unique nature of fine-tuning multiplicity in Tabular LLMs. [13, 79] argue for the necessity of measuring and reporting multiplicity to better inform predictions. Traditional methods to measure multiplicity in classical ML are impractical for LLMs due to the computational challenge

of retraining several fine-tuned models [13, 15, 101]. Our proposed measure, which requires only the given model and leverages the embedding space to inform multiplicity, addresses this issue. This approach reduces the complexity from retraining and inference to just inference, making it more feasible to apply in practice. Although, a large k (number of sampled points) may be needed for accurate stability estimation, it remains computationally more efficient than retraining multiple models. Compared to existing methods, our stability measure achieves a superior trade-off between runtime and correlation with evaluated multiplicity metrics, as shown in Figure 3.4. Our work provides practitioners with meaningful information about the multiplicity of predictions, which may lead them to carefully evaluate which predictions to trust and which to treat with caution. Our research has significant implications in several high-stakes applications, e.g., hiring, finance, education, etc., where inconsistent predictions can lead to distrust.

Broader Societal Impacts. The application of LLMs to tabular data, particularly in high-stakes domains such as finance and healthcare, presents both opportunities and risks [115]. Our work aims to address one of the critical challenges associated with these models: the instability introduced when fine-tuning large models on small datasets. This instability, manifested as overfitting and multiplicity, can undermine the reliability of model predictions in scenarios where stability is crucial. By measuring multiplicity, our work contributes to the responsible deployment of LLMs in domains where erroneous predictions can have severe consequences [94, 115]. Tabular data is central to these high-stakes domains but remains underexplored compared to text and vision [1]. Recent work emphasizes the need for reliable foundation models in this modality [64].

Our approach also supports *regulatory compliance* by enhancing transparency and accountability in automated decision-making systems. Quantifying prediction consistency aligns with regulations such as the General Data Protection Regulation (GDPR) [9] and upcoming AI legislation, which in-

creasingly demand explainable and reliable AI models [116]. While LLMs are more computationally expensive than traditional models, our method reduces the costs of assessing multiplicity. By avoiding repeated retraining, it enhances *cost efficiency* and *minimizes environmental impact*, lowering both energy consumption and carbon footprint [117].

Furthermore, observing the nature of fine-tuning multiplicity in Tabular LLMs pave the way for future research into model stability. It also *facilitates continual learning* by informing the robustness of a prediction to potential model updates in a dynamic environments where data constantly evolves [76, 118, 119]. Lastly, our work could play a role in mitigating *fairwashing risks* and *explanation bias* [77–79]. This transparency is crucial for maintaining ethical standards and trustworthiness in AI deployment [116].

Limitations. Our work provides a measure to assess fine-tuning multiplicity, but does not directly resolve this issue. Future research could focus on mitigation methods to ensure more consistent model predictions. A key constraint is the applicability to higher-dimensional datasets due to the limited context window size of current LLMs, though extending context windows is an active area of research [120, 121]. Additionally, our method's performance can be sensitive to hyperparameters, such as sample size and neighborhood radius; incorrect choices may lead to an inaccurate assessment of robustness. Our approach also assumes access to the embedding space, limiting its application to open-source models. Furthermore, the bound in Theorem 3.1 is not directly computable. Estimating these unknowns such as ϵ' could be a direction for future work. Despite these limitations, our measure serves as a crucial step toward understanding and quantifying fine-tuning multiplicity, laying the groundwork for future advancements.

Chapter 4: Improving Consistency in RAG Systems with Group Similarity Rewards

4.1 Introduction

LLMs are increasingly used in open-domain applications where users expect them to behave predictably, producing consistent outputs for semantically equivalent or paraphrased inputs. However, they frequently generate divergent responses to such variations, raising concerns about their reliability [122–124]. RAG systems are particularly prone to such inconsistencies [125]. These architectures combine a retriever and a generator: the retriever selects top-k documents from a large corpus, and the generator synthesizes a response conditioned on those documents [126]. Semantically similar queries can lead to different retrieved document sets or rankings, resulting in divergent outputs [125]. [127] also highlights a theoretical bottleneck in embedding-based retrieval, showing that the expressivity of top-k retrieval is fundamentally limited—underscoring the need for systems that are robust to retrieval inconsistencies. Furthermore, even when the evidence is fixed, the generator may still produce inconsistent responses due to the non-deterministic nature and phrasing sensitivity of LLMs [128].

This inconsistency is particularly problematic in high-stakes domains such as healthcare or legal settings, where RAG systems are commonly deployed [129]. Inconsistent outputs can erode trust, introduce liability risks, or even mislead users [122, 129]. For instance, a customer service RAG assistant may offer different instructions for "*How do I close my savings account?*" and "*What steps should I take to shut down my savings account?*" despite these queries being semantically equivalent [128].

In this work, we focus on *information consistency*—the requirement that outputs convey the same core content and information across paraphrased inputs (see motivational Figure 4.1). This contrasts with *lexical consistency*, which emphasizes word-level or structural similarity. While lexical consistency is easier to measure, it can penalize legitimate variation (e.g., use of synonyms or stylistic changes) and is insufficient in evaluating factual agreement. Crucially, the relationship between consistency and accuracy varies across QA tasks. In short-form QA, where answers are typically concise and factual, improving consistency often correlates with higher accuracy, models that are more consistent tend to be more correct. In contrast, for long-form QA tasks, where multiple valid answers may exist, consistency and accuracy become orthogonal dimensions: a model can be accurate yet inconsistent, or vice versa. Hence, in open-ended tasks, enforcing information consistency becomes a key desideratum alongside answer quality.

Given the practical importance of consistent outputs, we aim to address the following question: *How can we measure & improve the information consistency of RAG system outputs across semantically equivalent inputs, without compromising factual accuracy?* To tackle this, we introduce a new evaluation framework that decomposes consistency into retriever-level and generator-level components, and propose a reinforcement learning approach to optimize for consistency using group similarity rewards. Our contributions can be summarized as follows:

- **A Framework for Measuring Consistency in RAG Systems.** We present a principled framework to evaluate consistency in RAG systems by disentangling three components: retriever consistency (Jaccard overlap of documents), generator consistency (LLM outputs given fixed context), and end-to-end consistency. We instantiate this using lexical and LLM-Judge based similarity metrics, offering insights into where and how inconsistencies emerge (see Section 4.3.1).

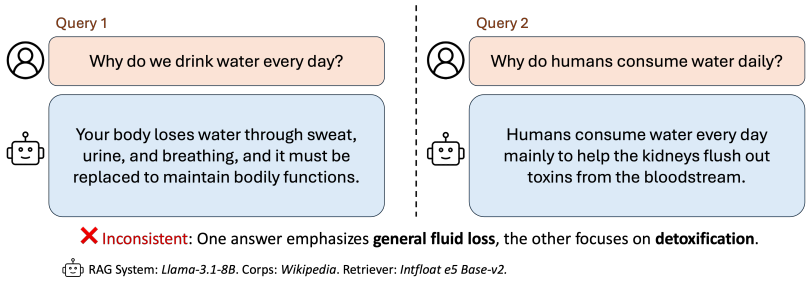


Figure 4.1: Two semantically equivalent queries lead to different outputs, despite both responses being factually correct. Such variation may be acceptable in many applications, but in certain high-stakes domains (e.g., healthcare, finance, legal) information consistency across semantically equivalent inputs may be required to ensure reliability, user trust, and compliance.

- **Con-RAG: Improving Consistency via Paraphrased Set GRPO.** To enhance consistency across semantically equivalent queries, we propose **P**araphrased **S**et GRPO, an RL approach that leverages multiple rollouts across a set of paraphrased inputs to assign *group similarity rewards*. This forms the core of our Information **Con**sistent **RAG** (Con-RAG) framework (see Figure 4.2). Due to complexity of computing the rewards, we introduce a relaxed approximation by subsampling paraphrases and rollouts, reducing the number of comparisons from quadratic to linear in the number of paraphrases. This allows us to train Con-RAG efficiently on large datasets while preserving reward fidelity.

- **Experments.** We conduct an extensive evaluation of Con-RAG across five QA benchmarks: short-form, multi-hop, and long-form QA tasks on Llama3.1 and Qwen2.5 model families. (see Figure 4.3). Our results show that Con-RAG significantly improves both end-to-end and generator consistency over a wide range of baselines, without degrading accuracy. In long-form QA tasks, Con-RAG improves both consistency and LLM-judged factual accuracy despite being trained in the absence of explicit ground-truth supervision.

4.2 Related Work

Consistency in Language Models. Consistency has emerged as a key concern for safety and reliability in high-stakes LLM deployment [122, 129]. Prior work has introduced various notions of consistency. Logical consistency refers to the ability of the model to make decisions without logical contradiction [130–133]. Factual consistency, often discussed as faithfulness or hallucination, considers whether model outputs contradict the source content [130, 134–136]. Self-consistency evaluates whether similar inputs yield stable explanations [137]. Nonlogical forms of consistency, such as moral consistency, assess coherence of values across contexts [138, 139]. Prediction consistency examines whether a model's predictions remain stable across multiple fine-tuned variants that achieve similar overall performance [71, 140]. Closest to our work is semantic consistency, which measures output stability under semantically equivalent inputs like paraphrases. This has been evaluated using datasets like ParaRel [123] and metrics such as BERTScore, entailment scores, and LLM judges [141–143]. Approaches to improve semantic consistenc include custom losses [123], knowledge distillation from consistent teachers [124], and synthetic data supervision [144]. We refer to a recent survey exploring current landscape, challenges, and future directions in consistency in LLMs [122].

Consistency in RAG Systems. RAG improves factual accuracy by conditioning outputs on retrieved evidence [145–147]. However, it introduces new sources of inconsistency due to retriever sensitivity and generator (LLM) variability. Despite growing use in high-stakes applications, information consistency in RAG remains underexplored, with the exception a few notable studies addressing robustness in retrieval or prompt-level variation [125, 148–150]. Our work aims to evaluate and improve information consistency in RAG, leveraging an RL-based optimization with group similarity rewards. Our approach builds on recent advances in RL for LLMs [151], particularly GRPO [152], which trains on

verifiable reward assignment across outputs. Our framework improves information consistency across semantically equivalent inputs without relying on ground-truth labels, unlike prior methods.

4.3 Main Contributions

In this section, we first define a framework to measure consistency in RAG systems by isolating retriever, generator, and end-to-end contributions (see Section 4.3.1), then introduce our Con-RAG method to improve consistency via group similarity rewards and its relaxation (see Section 4.3.2).

4.3.1 Measuring Consistency in RAG Systems

We consider a RAG system composed of a retriever R and a generator (LLM). Given a user query q, the system first retrieves a set of top-k documents from a corpus $\mathcal{D}$, and then generates an output $y = \text{LLM}(q, R(q))$ conditioned on these documents:$R(q) = \{d_1, \ldots, d_k\} \subset \mathcal{D}$. Let q_0 be a canonical input query, and let $\mathcal{P}(q_0) = \{p_1, p_2, \ldots, p_n\}$ denote a set of paraphrased or semantically equivalent inputs. Our goal is to assess the *output consistency* of the RAG system across this paraphrased set.

Retriever Consistency. Let $R(p_i)$ denote the set of documents retrieved for paraphrase p_i. We define retriever-level consistency as the average similarity between the document sets retrieved for all pairs of paraphrases. We use *Jaccard similarity* [153], which measures the ratio of the intersection to the union of two sets. This metric directly captures the overlap between retrieved evidence sets while normalizing for their total size. The overall retriever consistency is then the average across all unique paraphrase pairs: $\mathcal{C}_{\text{ret}}(q_0) = \frac{2}{n(n-1)} \sum_{i,j} \frac{|R(p_i) \cap R(p_j)|}{|R(p_i) \cup R(p_j)|}$.

End-to-End RAG Consistency. Let $y_i = \text{LLM}(p_i, R(p_i))$ denote the output of the RAG system for paraphrase p_i. End-to-end consistency measures alignment across outputs when the entire pipeline

is allowed to vary, each paraphrase p_i is passed to the retriever, which may return a different document set $R(p_i)$, and the generator then conditions on this evidence to produce y_i. Formally, we compute pairwise similarity across all outputs: $\mathcal{C}_{\text{gen}}(q_0) = \frac{1}{n(n-1)} \sum_{i \neq j} \text{sim}(y_i, y_j)$. This captures the overall stability of the RAG system under paraphrased inputs, reflecting the combined variability introduced by both retrieval and generation. The similarity function $\text{sim}(\cdot, \cdot)$ can be instantiated using various metrics, including lexical similarity (e.g., BLEU, ROUGE), embedding-based similarity (e.g., BERTScore), entailment-based scores from NLI models, or LLM-based judgments using a strong language model to assess consistency or contradiction between y_i and y_j.

Generator (LLM) Consistency. To isolate the generator's contribution, we can fix the retrieved documents across all paraphrases and measure similarity among the outputs, i.e., $y_i^{\text{fixed}} = \text{LLM}(p_i, R(q_0))$, and compute consistency over $\{y_1^{\text{fixed}}, \ldots, y_n^{\text{fixed}}\}$. This captures how consistently the LLM alone responds to semantically equivalent inputs when conditioned on identical evidence. Conceptually, this is closely related to prior work on consistency in standalone LLMs, where the focus is on ensuring paraphrase-invariant outputs under identical or similar prompts [122–124, 128].

4.3.2 Improving Consistency via Paraphrased Set GRPO

Given a RAG system comprise a retriever R and generator (LLM). A canonical query q_0 with paraphrases $\mathcal{P}(q_0) = \{p_1, \ldots, p_n\}$, our goal is to maximize output consistency without degrading factual accuracy. We propose Paraphrased Set GRPO, an RL algorithm that leverages GRPO's multiple rollouts across paraphrased inputs to assign group-level similarity rewards. Our objective is to directly optimize the generator so that outputs across semantically equivalent inputs are consistent.

Group Relative Policy Optimization. GRPO is RL optimization algorithm that estimates advantage

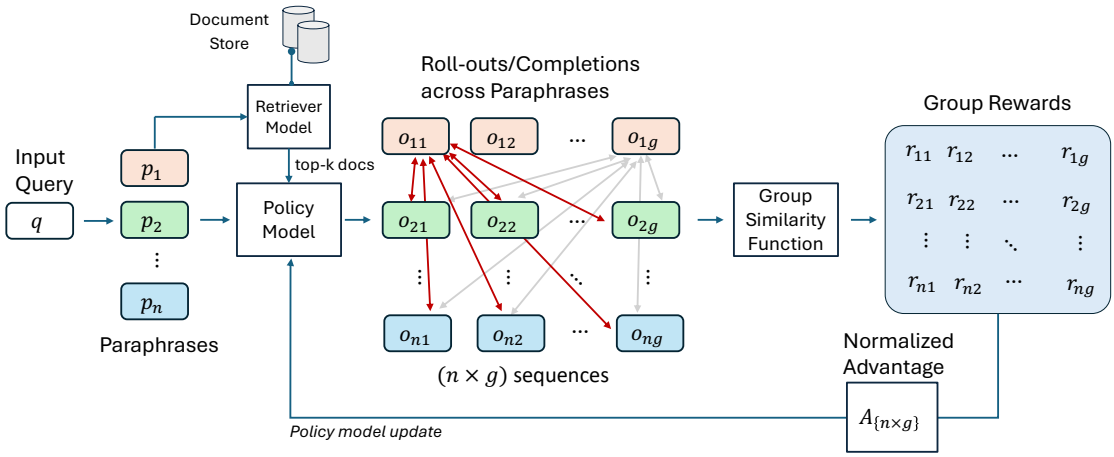


Figure 4.2: Overview of PS-GRPO and Information **Con**sistent **RAG** (Con-RAG) framework. A canonical query q is expanded into a set of paraphrases $\{p_1, \ldots, p_n\}$, each of which is passed through the policy LLM to generate g sampled rollouts. For every rollout o_{ij}, we compute a group similarity reward r_{ij} by averaging its similarity with outputs from other paraphrases of the same query (this produces an $n \times g$ reward matrix). Normalized advantages are then computed within each paraphrase set, and the policy model is updated.

through group-normalized rewards rather than using a critic model [152]. For a given query q, GRPO samples a group of g rollouts, i.e., multiple possible completions generated from the policy under stochastic decoding (such as temperature or nucleus sampling), denoted by $\{o_1, \ldots, o_g\}$, where each rollout is drawn as $o_i \sim \pi_\theta(\cdot \mid q)$. Each rollout receives a verifiable scalar reward $r_i = \text{Reward}(o_i \mid q)$, and the normalized advantage for each rollout is computed as $\hat{A}i = (r_i - \mu_q)/\sigma_q$, where μ_q and σ_q are the mean and standard deviation of rewards within the group. Let $y_{i,1:|o_i|}$ denote tokens of response o_i and $\rho_{i,t}=\frac{\pi_\theta(y_{i,t}|p,y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|p,y_{i,<t})}$. The policy is then optimized by maximizing the objective using these group-relative advantages, with an optional KL penalty to penalize deviation from the reference policy:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \frac{1}{g}\sum_{i=1}^{g}\sum_{t=1}^{|o_i|}\min\Big(\rho_{i,t}\hat{A}_i,\ \text{clip}(\rho_{i,t},\ 1-\epsilon,\ 1+\epsilon)\hat{A}_i\Big) - \beta\,\mathbb{D}_{\text{KL}}\big(\pi_\theta(\cdot \mid q)\,\big\|\,\pi_{\text{ref}}(\cdot \mid q)\big) \quad (4.1)$$

Group Similarity Rewards. PS-GRPO introduces a group-level objective that promotes consistent generation across semantically equivalent queries. It leverages the unique property of GRPO which

generates *extensive rollouts per query*. We aggregate all rollouts from all paraphrases into a group and compute similarity-based rewards *across the paraphrase set*, so each output is rewarded according to its similarity with outputs generated for the other paraphrases of the same canonical query. For each canonical query q_0 with paraphrases $\mathcal{P}(q_0) = \{p_1, \ldots, p_n\}$, the policy LLM π_θ generates g rollouts per paraphrase: $o_{ij} \sim \pi_\theta(\cdot \mid p_i, R(p_i)), i \in \{1, \ldots, n\},\ j \in \{1, \ldots, g\}$. Collect these into an $n \times g$ matrix $\{o_{ij}\}$ (total $n \times g$ rollouts). We assign each rollout o_{ij} a group similarity reward by averaging its similarity to all rollouts generated for the *other* paraphrases (also see Figure 4.2 for illustration):

$$r_{ij} \;=\; \frac{1}{(n-1)g} \sum_{\substack{u=1 \\ u \neq i}}^{n} \sum_{m=1}^{g} \mathrm{sim}\big(o_{ij},\, o_{um}\big), \tag{4.2}$$

where $\mathrm{sim}(\cdot, \cdot)$ is the agreement function. In practice, we instantiate sim using the BLEU metric, motivated by recent findings that BLEU serves as a strong proxy for reward models in aligning LLMs with human preferences [154]. As further validated in our ablation study (see Table 4.4), BLEU consistently outperformed alternative similarity metrics while remaining computationally efficient. Group-normalized advantages are then computed across each paraphrased rollout:$\hat{A}_{ij} = (r_{ij} - \mu_i)/\sigma_i$, with μ_i, σ_i the mean and standard deviation of rewards for rollouts for p_i. The policy is optimized with the GRPO objective using $\hat{A}_{ij}$ and (optionally) a KL penalty to a reference policy with weight β. If ground-truth answers are available (e.g., in short-form QA tasks), we extend the reward to improve consistency and accuracy. Specifically, for each rollout we define a combined weighted reward:

$$r_{ij}^{\text{final}} \;=\; \alpha\, r_{ij}^{\text{cons}} \;+\; \gamma\, \mathrm{Acc}(o_{ij}, y^\star), \tag{4.3}$$

where r_{ij}^{cons} is the group similarity reward, $y^\star$ is the ground-truth answer, and $\mathrm{Acc}(\cdot, \cdot)$ is measured

using token F1 score. Importantly, our method does not require ground truths to improve consistency: the accuracy reward term can be omitted, as demonstrated in our long-form QA experiments (see Section 4.4), where questions are open-ended and no single ground-truth answer exists.

Efficient Computation of Group Similarity Rewards for Scalable Training. Computing group similarity rewards can be expensive, especially in a training environment where rewards must be computed at every gradient step. This overhead can significantly slow down training. For each rollout o_{ij}, computing its reward requires comparing against all $(n-1)g$ rollouts from the other paraphrases. At the query level, with n paraphrases and g rollouts each, the naive total cost is $ng \times (n-1)g = n(n-1)g^2$ similarity computations. For example, with $n = 5$ and $g = 6$ amounts to 720 similarity comparisons for a single query. Exploiting symmetry (a similarity between o_{ij} and o_{um} need not be recomputed twice) reduces this to $\frac{1}{2}\, n(n-1)g^2$, but the cost still scales quadratically with n and g. To make training feasible, we introduce a *relaxed group similarity reward*. Instead of averaging over all cross-paraphrase comparisons, for each rollout o_{ij} we subsample κ paraphrases $K \subset \{1, \ldots, n\} \setminus \{i\}$ and s rollouts per chosen paraphrase, and approximate: $\tilde{r}_{ij} = \frac{1}{\kappa s} \sum_{u \in K} \sum_{m \in S_k} \mathrm{sim}(o_{ij}, o_{um})$, which is an unbiased estimator under uniform sampling. This reduces the per-query cost from $O(n(n-1)g^2)$ to $O(ng\kappa s)$, if $\kappa \ll n-1$ and $s \ll g$. In practice, this approximation preserves the training signal for cross-paraphrase consistency while keeping the reward computation tractable.

4.4 Experiments

In this section, we describe our experimental setup to evaluate the effectiveness of Con-RAG across diverse QA tasks, outlining our datasets, paraphrase generation, consistency metrics, training details, and comparisons with competitive baselines.

Table 4.1: Disentangling sources of inconsistency in RAG systems (LLaMA-3.1-8B). Retriever consistency is low across datasets, suggesting that paraphrased queries often retrieve non-overlapping documents. This introduces context variability that is reflected in the end-to-end consistency scores. Fixing retrieval improves consistency, but variation remains, revealing the generator's sensitivity to input phrasing even with identical evidence. We present accuracy values in Table 4.5 (also see Table C.2 for Qwen-2.5-3B).

Dataset	**End-to-End Consistency**		**Generator (LLM) Consistency**		**Retriever Consistency**
	Lexical	LLM-Judge	Lexical	LLM-Judge	Jaccard Overlap
TriviaQA	53.0	77.8	67.3	88.5	32.5
HotpotQA	42.5	62.5	53.7	71.9	46.0
2Wiki	38.5	65.5	48.4	76.4	52.4
MuSiQue	27.9	48.2	44.4	69.7	36.6
Eli5	8.56	62.8	15.1	74.2	27.1

Datasets. We evaluate our approach across three types of question answering (QA) tasks: Short-form QA tasks: TriviaQA [155] and HotpotQA [156], both requiring concise fact-based answers. Multi-hop QA tasks: 2WikiMultiHopQA [157] and MuSiQue [158], which involve reasoning over multiple pieces of evidence. Long-form QA task: ELI5 [159], where answers are open-ended and typically span multiple sentences. None of these datasets provide paraphrased versions of the input questions. To evaluate consistency, we synthetically generate paraphrases for each query.

Generating semantically equivalent queries. For each query q_0, we use LLaMA-3.1-70B to generate n paraphrases $\mathcal{P}(q_0) = \{p_1, \ldots, p_n\}$. To ensure answerability, we provide the ground truth answer as part of the prompt and instruct the model to generate paraphrases that preserve the exact meaning such that each paraphrase can be answered in the same way. This allows us to simulate semantically equivalent inputs without altering the expected outputs (see prompt in Appendix C.2).

Setup. Our RAG system consists of a LLaMA-3.1-8B and Qwen-2.5-3B model serving as the generator, and a dense retriever built on top of the intfloat/e5-base-v2 embedding model [160]. We use KILT Wikipedia snapshot [161] as our document corpus, where each article is segmented into chunks of 512 tokens before embedding. All embeddings are indexed using FAISS for efficient retrieval. At

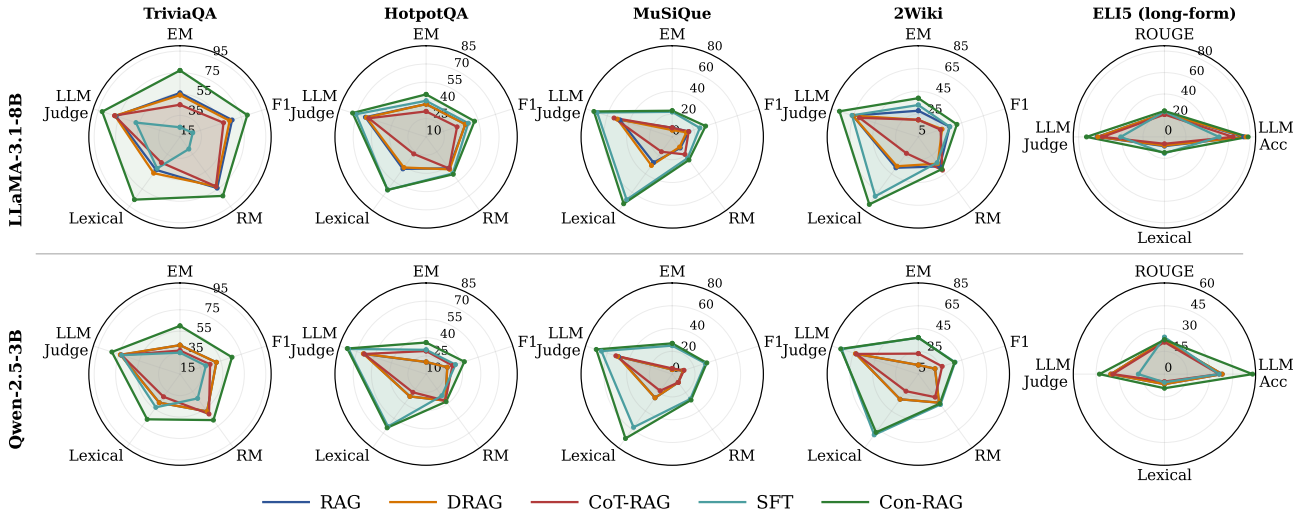


Figure 4.3: Comparison between Con-RAG and baselines across accuracy and consistency dimensions on LLaMA-3.1-8B and Qwen-2.5-3B. Each plot summarizes performance on a single dataset using accuracy measures (Exact Match, token F1, Relaxed Match) and end-to-end information consistency (measured lexically and via LLM-judge). Con-RAG consistently outperforms prior methods across models, achieving both higher factual accuracy and more consistent responses across paraphrased inputs (see Table 4.2 for full numerical results).

inference time, the retriever selects the top-$k = 5$ documents per query, which are then appended to the prompt for generation. To isolate effects from sampling inconsistencies, we use deterministic decoding throughout all experiments.

Evaluating Consistency in RAG Systems. We evaluate performance along two dimensions: accuracy and consistency. For short-form and multi-hop QA datasets, accuracy is measured using: (i) Exact Match (EM), (ii) token F1 score, and (iii) Relaxed Match (RM), which considers an answer correct if the ground truth answer appears anywhere in the output. For long-form QA (e.g., ELI5), where answers are open-ended and may be phrased in diverse ways, EM/F1/RM are too restrictive. Instead, we evaluate accuracy using: (i) ROUGE, to capture content overlap with reference answers, and (ii) LLM-judge accuracy, where a strong model (LLaMA 3.3 70B) assesses whether the generated answer is factually correct. Consistency is evaluated at three levels (disentangling contributions from the retriever and generator): (i) End-to-end consistency, where each paraphrase retrieves its own documents

and we compute agreement between outputs (via BLEU for lexical consistency and an LLM judge for information consistency); (ii) Generator consistency, where retrieval is fixed across paraphrases and agreements cross outputs are measured; (iii) Retriever consistency, defined as the average Jaccard overlap between retrieved document sets across paraphrases (see Section 4.3.1). We use paraphrase size $n = 5$ for evaluations (see Appendix C.2 for prompts templates).

We summarize consistency results across the datasets in Table 4.1. We observe that the retriever consistency is relatively low across the datasets, indicating that paraphrases often retrieve non-overlapping sets of documents, a key source of downstream inconsistency. This is reflected in the end-to-end consistency scores, which shows that these small changes in query phrasing can result in different answers, due to shifts in both retrieved context and model generation. To isolate the generator's contribution, we also evaluate generator consistency under fixed retrieval (i.e., same documents across paraphrases). While consistency scores improve, substantial variability still remain, showing that even with identical evidence, the generator (LLM) exhibits sensitivity to input phrasing.

We report accuracy for original queries, paraphrased queries, and paraphrased queries with fixed documents in Table C.1. Across these settings, accuracy remains relatively stable, with only minor fluctuations, suggesting that paraphrasing and retrieval shifts have limited impact on final answer correctness on average.

Con-RAG Training Setup. We train Con-RAG with BLEU as similarity function for computing group similarity rewards. For short-form and multi-hop QA tasks, we use unigram BLEU (ngram=1) and bigram BLEU (ngram=2) for long-form QA tasks to account for more contextual similarity across longer answers. For short-form QA tasks, where ground-truth answers are available, we augment the similarity reward with an accuracy reward based on token F1 score, which we found to be more stable than other accuracy metrics. The final reward is computed using a weighted sum as defined in

Table 4.2: Comparison between Con-RAG vs. Baselines (Short-form QA Tasks) (LLaMA-3.1-8B). Lexical consistency measured via BLEU score while and information consistency measured using an LLM-judge. Con-RAG is trained with a group similarity reward plus an accuracy reward (no KL), and consistently yields higher end-to-end and generator-only consistency while also improving accuracy over original queries (see radar plot illustration in Figure 4.3). Refer to Table C.4 for results on Qwen-2.5-3B model.

		Accuracy (%)			**End-to-End Consistency (%)**		**Generator (LLM) Consistency (%)**	
Dataset	**Method**	EM	F1	RM	Lexical	Inform.	Lexical	Inform.
TriviaQA	RAG	56.0	66.1	74.0	53.0	77.8	67.3	88.5
	DRAG	54.0	63.7	72.0	56.8	78.7	68.2	88.2
	CoT-RAG	45.0	57.7	72.0	44.6	79.2	57.7	85.0
	SFT	24.0	27.5	29.0	51.3	58.2	77.8	81.2
	Con-RAG	**77.0**	**81.0**	**83.0**	**87.3**	**91.3**	**91.2**	**93.0**
HotpotQA	RAG	37.0	44.1	42.0	42.5	62.5	53.7	71.9
	DRAG	37.0	43.8	43.0	41.1	61.6	50.5	73.1
	CoT-RAG	31.0	36.8	42.0	27.3	59.6	36.1	68.9
	SFT	39.7	46.5	47.2	**63.9**	70.5	72.2	78.5
	Con-RAG	**45.0**	**51.9**	**48.0**	**63.9**	**73.6**	**80.9**	**88.2**
MuSiQue	RAG	8.0	15.3	12.0	27.9	48.2	44.4	69.7
	DRAG	6.0	13.1	11.0	31.0	50.7	42.9	70.0
	CoT-RAG	8.0	15.2	19.0	16.1	53.7	29.2	67.7
	SFT	22.0	25.5	23.0	68.1	69.3	77.8	79.8
	Con-RAG	**23.0**	**30.8**	**25.0**	**72.5**	**72.3**	**91.4**	**92.7**
2Wiki	RAG	28.0	33.9	37.0	38.5	65.5	48.4	76.4
	DRAG	20.0	26.9	34.0	36.8	65.5	49.3	76.1
	CoT-RAG	20.0	25.5	41.0	22.8	59.3	29.9	67.8
	SFT	33.0	34.0	33.0	69.4	66.2	84.4	83.3
	Con-RAG	**39.0**	**40.6**	**40.0**	**78.2**	**77.8**	**94.1**	**95.5**

Eq. 4.3, with equal weights $(\alpha, \gamma = 1)$ for both consistency and accuracy. We set the KL regularization coefficient $\beta = 0.0$ for these tasks, following recent findings [162] suggesting that GRPO performs effectively without explicit KL penalties. In contrast, for long-form QA (ELI5), where questions are open-ended and multiple valid answers may exist, we exclude the accuracy reward and optimize solely for consistency using the group similarity reward. To prevent reward hacking in the absence of ground-truth supervision, we apply a small KL penalty with $\beta = 0.05$ to regularize the policy against a reference model.

Table 4.3: Comparison between Con-RAG vs Baselines (Long-form QA Task). Con-RAG is trained using only the group similarity rewards with a small KL regularizer (no accuracy supervision). Despite no ground-truth, it achieves the best end-to-end and generator consistency and also improves answer quality over baselines, whereas SFT on reference answers underperforms in this open-ended setting.

		Accuracy (%)		**End-to-End Consistency (%)**		**Generator (LLM) Consistency (%)**	
Dataset	**Method**	ROUGE	LLM-Acc	Lexical	Inform.	Lexical	Inform.
ELI5	RAG	21.9	74.0	8.6	62.8	15.1	74.2
	DRAG	22.0	76.0	8.0	62.2	15.0	72.5
	CoT-RAG	20.9	64.0	6.4	57.8	10.3	71.0
	SFT	23.5	51.0	**15.3**	40.8	16.6	41.7
	Con-RAG	**24.2**	**78.0**	14.6	**72.7**	**21.7**	**80.8**

We use $n = 6$ paraphrases per canonical query and $g = 4$ rollouts per paraphrase. To make training scalable, we apply the relaxed approximation described in Section 4.3.2 to estimate group similarity rewards. Specifically, we subsample $\kappa = 3$ paraphrases and $s = 1$ rollout per selected paraphrase when computing similarity, which significantly reduces the number of comparisons with minimal impact on reward quality. We perform full model fine-tuning using the AdamW optimizer with a learning rate of `1e-6`. All training is conducted on LLaMA-3.1-8B and Qwen-2.5-3B.

Baselines. We compare Con-RAG against diverse baselines representative of current RAG systems: (i) **RAG**: A standard RAG setup where the top-k retrieved documents are appended to the prompt and passed directly to the generator for answer prediction. (ii) **DRAG** (Demonstrated RAG) [163]: An inference-time scaling method that leverages few-shot demonstrations to improve performance. (iii) **CoT-RAG** (Chain-of-Thought RAG) [164]: Extends standard RAG by prompting the generator to produce intermediate reasoning steps before outputting a final answer, improving multi-hop and compositional question answering. (iv) **SFT** (Supervised Fine-Tuning) [83]: We fine-tune the generator on paraphrased queries paired with their ground-truths. For long-form QA, where answers are free-form, we fine-tune on the available reference responses. (v) **Con-RAG** (ours): Our proposed method

Table 4.4: Effect of Reward Similarity Metric on Con-RAG (`ELI5-Qwen-2.5-3B`). We vary the similarity function used in the group reward to study its impact on information consistency. Lower-order BLEU emphasizes word choice and local fluency, aligning better with the goal of preserving core information across paraphrases. In contrast, higher-order BLEU and Exact Match enforce stricter surface-level or sentence-level overlap, which can penalize valid rephrasings. BLEU-2 yields the best consistency and accuracy, indicating that rewarding semantic adequacy is better aligned with information consistency.

Reward Metric	**Accuracy (%)**		**End-to-End Cons. (%)**		**Generator Cons. (%)**	
	ROUGE	LLM-Acc	Lexical	LLM-Judge	Lexical	LLM-Judge
BLEU-1	**22.6**	54.0	6.9	38.2	14.8	**69.8**
BLEU-2	22.5	**58.0**	**9.2**	**42.0**	**17.8**	67.5
BLEU-3	22.4	49.0	6.7	36.3	14.8	66.0
BLEU-4	22.2	50.0	6.4	36.2	14.2	66.5
ROUGE-L	22.1	46.0	6.1	35.2	13.6	65.2
Exact Match	22.1	49.0	6.6	37.7	14.4	66.2

that leverages group similarity rewards to improve consistency (see Section 4.3.2). All baselines are evaluated using the same retriever, generator, and document corpus to ensure fair comparison.

Results and Analysis. We present our results across short-form and long-form QA tasks in Figure 4.3 and Tables 4.2. To show that consistency improvements do not come at the cost of answer quality, we report accuracy metrics on the original queries, avoiding generic but consistent outputs. Our results demonstrate the following key observations:

Con-RAG improves both consistency and accuracy in short-form QA. Across all short-form datasets, Con-RAG achieves significant gains in both end-to-end and generator-only consistency. For instance, on TriviaQA, end-to-end consistency (lexical/information) improves from $53.0/77.8$ to $87.3/91.3$, while generator consistency reaches $91.2/93.0$. Notably, these improvements are not achieved at the expense of accuracy. Con-RAG also achieves the highest EM, F1, and RM scores across all datasets. This indicates that optimizing consistency can also enhance model robustness, likely due to the implicit data augmentation effect of training across paraphrased inputs. Other baselines DRAG and CoT-RAG provide only modest consistency improvements and fail to match Con-RAG across metrics.

In Long-form QA, Con-RAG also boosts accuracy without ground-truth supervision. Results on ELI5 (see Table 4.3) are particularly interesting: even though Con-RAG is trained without any explicit ground truth (or accuracy signal), it improves both consistency and accuracy over all baselines. Compared to RAG, Con-RAG increases lexical and information consistency while also achieving higher ROUGE and LLM-judged accuracy. In contrast, SFT trained on reference answers performs poorly on ELI5, especially in terms of LLM-judge accuracy, highlighting the limitations of rigid supervision in open-ended QA, where many valid responses exist. This underscores the strength of Con-RAG in open-ended tasks, which does not rely on a single reference output.

Ablation Studies. To analyze design choices in Con-RAG, we run focused ablations on a lighter generator, Qwen-2.5-3B for fast, controlled sweeps. *1.) Varying similarity function used in the group reward.* We replace BLEU in the group similarity reward with alternative choices and measure resulting consistency/accuracy. We consider: BLEU-n ($n \in \{1, 2, 3, 4\}$), ROUGE-L, Exact Match (results are summarized in Table 4.4). *2.) Varying short-form accuracy reward metrics.* On short-form QA, we study the effect of different reward signals on accuracy and consistency by conducting ablations with: (i) consistency term only training, (ii) accuracy term only training, and (iii) joint training with consistency plus accuracy. For the accuracy component, we compare token F1 (ours), EM, and RM (see Table 4.5). *3.) Effect of LLM decoding temperature on consistency and accuracy.* We evaluate how inference-time stochasticity impacts consistency and accuracy by sweeping temperature values $T \in \{0.0, 0.5, 1.0, 2.0\}$ during decoding (see results in Table C.6).

Table 4.5: Effect of Accuracy Reward Variant on Con-RAG (`TriviaQA-Qwen-2.5-3B`). We compare consistency-only training, accuracy-only training, and joint training with consistency plus various accuracy metrics. The best performance is achieved when combining consistency with the token F1 reward, which yields the highest accuracy and consistency values.

Reward Variant	α	γ	**Accuracy (%)**			**End-to-End Cons. (%)**		**Generator Cons. (%)**	
			EM	F1	RM	Lexical	LLM-Judge	Lexical	LLM-Judge
Consistency only	1.0	0.0	51.5	53.2	59.0	59.9	79.0	78.7	88.0
Accuracy only (F1)	0.0	1.0	54.0	56.0	60.4	52.0	75.0	62.0	84.1
Consistency + EM	1.0	1.0	56.2	63.5	65.0	61.5	80.2	76.0	88.4
Consistency + RM	1.0	1.0	57.0	64.0	66.0	62.3	80.5	77.0	88.5
Consistency + F1	1.0	1.0	**60.0**	**66.0**	**68.0**	**67.1**	**81.8**	**80.5**	**89.5**

4.5 Discussion

While Con-RAG achieves strong improvements in both generator and end-to-end consistency, several important directions remain as next steps. (1) *Beyond Lexical Rewards for Information Consistency:* In this work, we use lexical similarity metrics (e.g., BLEU) as a proxy to enforce information consistency. While effective, such metrics emphasize surface-level alignment and penalize variations in wording, even when the underlying information remains unchanged. In practice, we may allow use of synonyms or outputs expressed differently, as long as they convey the same core content. A key next step is to search for a signal that would directly optimize for information-level consistency without enforcing lexical similarity between outputs. LLM as a judge seems promising, however, such a signal introduces a tension between weak vs. strong supervision [165]. Ideally, we seek lightweight, automatic signals that can still guide the model toward consistent output (leveraging entailment-based rewards, BERTScore, etc.). (2) *Joint Retriever and Generator Optimization:* Con-RAG substantially improves generator consistency, yet end-to-end consistency still lags behind, mainly due to variation in retrieved documents across paraphrased queries. This inconsistency in retrieval results in different contexts being provided to the generator. To address this, a promising next step is to jointly optimize

the retriever and generator. By rewarding the retriever to return similar documents for semantically equivalent queries, and simultaneously training the generator for consistency, the system can learn to retrieve relevant evidence that best helps answer the question accurately, potentially further improving both consistency and accuracy [147]. By introducing a principled way to measure RAG consistency and a scalable method to improve it, we move toward more reliable and user-aligned RAG systems.

Chapter 5: Few-Shot Distillation of LLMs With Counterfactual Explanations

5.1 Introduction

LLMs have demonstrated state-of-the-art performance across a broad spectrum of tasks [166–168]. However, as the size of LLMs grow, so does the associated computational burden, making them difficult to deploy in resource-constrained environments, e.g., mobile phones, edge devices, and embedded systems [169]. The challenge, therefore, lies in making large models more efficient and accessible without sacrificing performance. To this end, knowledge distillation (KD) (initially proposed in [170]; see surveys [171–173]) has emerged as a powerful technique for model compression, enabling smaller student models to mimic the performance of a larger teacher model. In the context of LLMs, KD plays a central role in transferring the broad capabilities such as natural language understanding [174], reasoning [175], instruction following [176], etc, onto smaller models.

While LLMs are trained for a broad range of tasks, we may often want a smaller, task-specific language model when full task coverage is not required. To support this, *task-aware knowledge distillation* [177, 178] has been proposed to selectively transfer task-relevant knowledge from teacher to student language models. While effective, these methods typically assume access to large datasets [179]. However, in many real-world applications, the amount of data available is often limited [180–183]. Despite advances on algorithmic strategies for task-aware KD in LLMs [179], the problem of data selection for KD has received limited interest, particularly in few-shot settings. In this work, we study

few-shot and task-aware knowledge distillation for LLMs, where student models are distilled from teacher models using a small number of samples labeled for a task (also called shots). Few-shot task-aware distillation remains underexplored for LLMs. In classical ML, few-shot training has poor generalization [184], and thus causes ineffective distillation due to insufficient task coverage [185, 186]. However, few-shot distillation holds potential for LLMs because they are pretrained on a large corpora, also drawing inspiration from the prior success of few-shot learning [187].

In this work, we propose a few-shot task-aware knowledge distillation by systematically integrating a type of post hoc explainability technique called counterfactual explanations (CFE) [25, 188, 189]. CFEs are inputs that flip the output prediction of a model with minimum input perturbations. We find that CFEs can act as *knowledge probes*, helping the students mimic the teacher's decision boundaries more effectively than standard data. Our work bridges explainability and model compression by turning explanations into actionable training signals, guiding the student into learning the teacher's decision-making process more effectively. This results in more faithful knowledge transfer even with very limited data. Our contributions can be summarized as follows:

- **A counterfactual explanation-based strategy for few-shot distillation.** We propose a novel framework CoD, short for **Co**unterfactual-explanation-infused **D**istillation, for task-aware knowledge distillation under few-shot regimes. By enriching the few-shot training set with CFEs, we improve the student's ability to mimic the fine-grained details of the teacher's decision boundary with fewer labeled examples. We validate this intuition through a synthetic experiment on the 2D moons dataset, showing that CFE-infused distillation better replicates the teacher's decision surface compared to using standard few-shot samples (see Fig. 5.3).

- **Theoretical guarantees motivating the role of CFEs in distillation.** We provide theoretical guar-

antees that serve as motivation for our approach, from both statistical and geometric perspectives. First, in a logistic regression setting, we show that CFEs improve parameter estimation by maximizing the Fisher Information (see Def. 5.2 & Thm. 5.1). Our proof specifically leverages the fact that the CFEs lie quite close to the decision boundary to show that they reduce the expected estimation error of the student model compared to standard distillation. Next, moving beyond statistical guarantees and linear models, we also provide a geometric analysis for non-linear models, establishing that if a student matches the teacher's predictions on the original data and their counterfactual pairs, then their decision boundaries will remain close: this is quantified by a provably small *Hausdorff distance*, a formal measure of distance between two subsets within a space (see Def. 5.3 & Thm. 5.2).

- **Empirical validation.** We evaluate CoD across six benchmark datasets using DeBERTa-v3 [190] and Qwen2.5 [167] model families. We compare against strong baselines for task-aware knowledge distillation including standard Knowledge Distillation (KD) [170], Layer-wise Distillation (LWD) [178], and Task-aware layer-wise Distillation (TED) [177] under various few-shot settings (k = 8, 16, 32, 64, 128, and 512). Our results demonstrate that CoD consistently outperforms baselines in few-shot regimes, with particularly significant improvements in extremely data-scarce scenarios ($k \leq 64$). Notably, COD only uses *half of the original labeled samples used by the baselines* (i.e., $k/2$ original infused with their corresponding $k/2$ CFEs, leading to k shots), and still gives improved performance. For instance, with $k = 8$ samples on IMDB dataset, LWD + CoD improves over standard LWD by more than 10 points (86.1% vs. 76.0%).

5.2 Related Works

Knowledge distillation has emerged as a powerful framework for model compression [170]. While early works focused on transferring soft labels via output logits [191], subsequent advances explored richer supervision signals such as intermediate feature alignment [178, 192–194]. As LLMs grow in size and inference cost [195, 196], distillation has become increasingly important for transferring capabilities into smaller models [171–173, 179]. More recently, task-aware knowledge distillation for LLMs has gained traction, aiming to selectively distill knowledge relevant to a specific downstream task [177, 197]. Despite these algorithmic innovations [177], there has been relatively little focus on data selection for distillation, particularly in few-shot settings. Most prior works assume ample training data, leaving few-shot knowledge distillation largely underexplored. While some works [181–183, 185] have studied distillation in classical ML under low-data regimes, they do not address the challenges specific to distilling LLMs. In this work, we establish the paradigm of few-shot distillation in LLMs by integrating explainable data selection. Our work is broadly aligned with the spirit of data-efficient ML, which aims to improve performance under limited supervision [198–200].

Counterfactual explanations (CFEs) [25, 39–41, 188, 189, 201] have been widely studied in classical ML, particularly in high-stakes applications such as finance, healthcare, and law, where they often provide algorithmic recourse to guide users toward desired outcomes [189, 202]. A closely related work [201] leverages CFEs for model reconstruction by deriving theoretical relationships between reconstruction error and the number of counterfactual queries using polytope theory [203]. In the natural language domain, some methods have been proposed to generate semantically valid CFEs using either token-level perturbations [204] or controlled generation with language models [204–206], but they have not been integrated for knowledge distillation. Another line of work is counterfactual

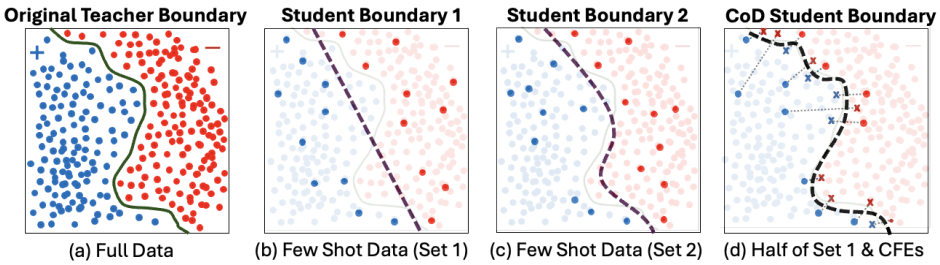


Figure 5.1: Intuition behind our approach: *(a)* Teacher trained on the full dataset with true decision boundary. *(b–c)* With few-shot supervision, many classifiers can fit the sparse points; the resulting student boundaries (dashed lines) can vary and do not always align with the teacher's boundary (unfaithful distillation). *(d)* Pairing each point with its CFE (×, linked to originals) during distillation makes the student match the teacher's soft predictions at these points. CFEs act as boundary-near pegs that clamp the student to the teacher's decision surface, producing a more faithful distillation even under few-shot budgets.

reasoning in causal inference, where the goal is to estimate the effect of interventions under a structural causal model [207], which is different from our objectives. These counterfactual data have been used to address the issue of spurious patterns in NLP tasks [208, 209], improve generalization [210, 211], and enhance performance on out-of-distribution data [212, 213]. In contrast, our work studies the role of CFE infusion in few-shot task-aware knowledge distillation, leveraging the teacher's signal to more effectively mimic the teacher's decision boundary in few-shot data settings.

5.3 Preliminaries

LLMs are highly effective for natural language processing. Built upon the transformer architecture [214], LLMs consist of multiple stacked layers, each containing a multi-head self-attention mechanism followed by a position-wise feed-forward neural network. Let $g(\cdot;\theta)$ denote a transformer-based model parameterized by θ. The model takes an input sequence $\mathbf{x} \in \mathcal{X}$ where $\mathcal{X}$ is the input space. The model output is a probability distribution over the vocabulary space, but for task-aware settings such as sentiment analysis, it is a probability distribution over C class labels, i.e., $g : \mathcal{X} \to [0,1]^C$. The loss function is defined as: $\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{X}}[\ell(g(\mathbf{x};\theta))]$, where ℓ denotes the task-specific loss, such as

cross-entropy for classification tasks or causal language modeling loss for generative models.

Knowledge Distillation (KD). KD is a technique that transfers knowledge from a large, pre-trained teacher model to a smaller, student model [215]. Let $g_t(\cdot;\theta_t)$ be the teacher model with parameters θ_t and $g_s(\cdot;\theta_s)$ be the student model with parameters θ_s. The teacher model $g_t(\cdot;\theta_t)$ provides soft labels to assist in training the student model $g_s(\cdot;\theta_s)$. The student is trained using a loss function that is a combination of the task-specific loss and the distillation loss as follows: $\min_{\theta_s} \mathcal{L}(\theta_s) + \alpha\mathcal{L}_{\text{KD}}(\theta_t, \theta_s)$. Here, $\mathcal{L}(\theta_s)$ is the task-specific loss, e.g., the cross-entropy loss between the student's outputs and true-labels, and $\mathcal{L}_{\text{KD}}(\theta_t, \theta_s) = \mathbb{E}_{\mathbf{x}\sim\mathcal{X}}[d(g_t(\mathbf{x};\theta_t), g_s(\mathbf{x};\theta_s))]$ is the distillation loss which captures the distance between the outputs of the teacher and student. Typically, the distance is computed using the Kullback-Leibler (KL) divergence, i.e., $\text{KL}(g_t(\mathbf{x};\theta_t) \,\|\, g_s(\mathbf{x};\theta_s)) = \sum_{c=1}^{C} g_t^{(c)}(\mathbf{x};\theta_t) \log \frac{g_t^{(c)}(\mathbf{x};\theta_t)}{g_s^{(c)}(\mathbf{x};\theta_s)}$, where the superscript (c) is for the assigned probability for class c by each model.

Layer-Wise Distillation (LWD). In large transformer-based models, the teacher's outputs may not fully capture the knowledge embedded in intermediate layers. Beyond matching final outputs, one can also align the intermediate features of the teacher and student [178]. At a few selected layers, the teacher's hidden activations h_t^l and the student's activations h_s^l (optionally projected into the same dimension) are computed and their difference is also penalized using a mean-squared-error loss [178]. The student is trained using a loss as follows:

$$\min_{\theta_s} \quad \mathcal{L}(\theta_s) \;+\; \alpha\, \mathcal{L}_{\text{KD}}(\theta_t, \theta_s) \;+\; \beta\, \mathcal{L}_{\text{LWD}}(\theta_t, \theta_s) \tag{5.1}$$

Here, $\mathcal{L}_{\text{LWD}}(\theta_t, \theta_s)$ is the additional layer-wise alignment term added alongside the task-specific loss and distillation loss, e.g., $\mathbb{E}_{\mathbf{x}\sim\mathcal{X}}[\sum_{l\in\mathcal{I}} \| h_t^l - h_s^l \|_2^2]$ where $\{h_t^l, h_s^l\}_{l\in\mathcal{I}}$ are the teacher and student activations for a given input $\mathbf{x}$ over a set $\mathcal{I}$ of layers, and $\alpha, \beta \geq 0$ balance the three objectives.

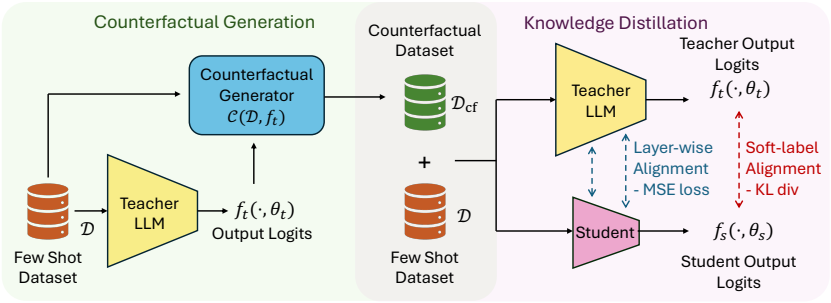


Figure 5.2: Overview of our framework: **Co**unterfactual Explanation-Infused **D**istillation (CoD).

Counterfactual Explanations (CFEs). Given a model's decision on an input $\mathbf{x}$, a CFE [25, 188, 189] finds the minimal modification $\mathbf{x}'$ such that the model's output changes in a desired way. These explanations help interpret model decisions and provide actionable guidance to users to flip the prediction. In our context, we look into CFEs in the NLP domain where the inputs are token sequences. A counterfactual in this setting is a minimally perturbed sentence that causes the teacher LLM's prediction to flip. For instance, given the sentence *I loved the movie*, labeled as positive sentiment, a CFE would be *I hated the movie*, a semantically similar but sentiment-flipped variant.

Our Problem Setting. We consider a binary classification setting where the teacher model will be denoted as $f_t : \mathcal{X} \to [0, 1]$. The input space $\mathcal{X} \subseteq \mathbb{R}^{n \times d}$, with n being the sequence length and d is the model dimension, after the entire input sequence has already been passed through the tokenizer and embedding layers of the LLM. The teacher model $f_t(\mathbf{x})$ gives the class-1 probability output of the model for input $\mathbf{x}$, i.e., $f_t(\mathbf{x}) := g_t^{(1)}(\mathbf{x}; \theta_t)$, where the superscript (1) is for the assigned probability for class 1. The final predicted class is given by $\hat{f}_t(\mathbf{x}) = \mathbb{I}\left[f_t(\mathbf{x}) \geq 0.5\right] \in \{0, 1\}$.

Definition 5.1 (Closest CFE $\mathcal{C}(\mathbf{x}, f_t)$)**.** *Given $\mathbf{x} \in \mathbb{R}^{n \times d}$ such that $f_t(\mathbf{x}) < 0.5$, the closest CFE is a point $\mathbf{x}' \in \mathbb{R}^{n \times d}$ with opposite prediction that minimizes the Frobenius-norm $\|\mathbf{x} - \mathbf{x}'\|_F$:*

$$\mathcal{C}(\mathbf{x}, f_t) = \underset{\mathbf{x}' \in \mathbb{R}^{n \times d}}{\arg\min} \|\mathbf{x} - \mathbf{x}'\|_F \text{ such that } f_t(\mathbf{x}') \geq 0.5. \tag{5.2}$$

Definition 5.1 naturally extends to multiclass settings, where a CFE can be defined as the minimum perturbation that changes the predicted class to any other target class.

Remark 5.1 (Data Manifold Counterfactual Explanations)**.** *In practice, unconstrained counterfactuals may lead to unrealistic or out-of-distribution examples. To address this, we can constrain* $\mathbf{x}'$ *to lie within the data manifold* $\mathcal{X}' \subseteq \mathbb{R}^{n \times d}$*, ensuring that generated counterfactuals remain semantically plausible. These* data-manifold counterfactuals *preserve natural language structure. In our work, we use a hybrid generation strategy that combines LLM-based prompting with teacher model feedback to generate such data-manifold CFEs. Further details are provided later in Section 5.4.*

Given a training data budget k (few-shots) and a teacher model f_t, our goal is to distill a smaller student model $f_s : \mathcal{X} \to [0, 1]$ with high-performance at a specific task by leveraging CFEs.

5.4 Main Contributions

We begin with an experiment on 2D synthetic data that demonstrates how CFEs help student models mimic the teacher's decision boundary more effectively than standard data. Next, we provide theoretical results motivating our approach from both statistical and geometric perspectives. Finally, we describe our CFE generation pipeline for natural language inputs, which leverages LLMs to produce semantically plausible CFEs, leading to our proposed framework CoD.

5.4.1 Synthetic Dataset Experiments to Illustrate the Role of CFE in Distillation

We conduct experiments on the 2D moons dataset [216] and show that infusing few-shot data with CFEs significantly improves student-teacher alignment in distillation (see Figure 5.3). We train a *teacher* model—a two-layer neural network with architecture $[2 \to 64 \to 64 \to 2]$ on the full dataset.

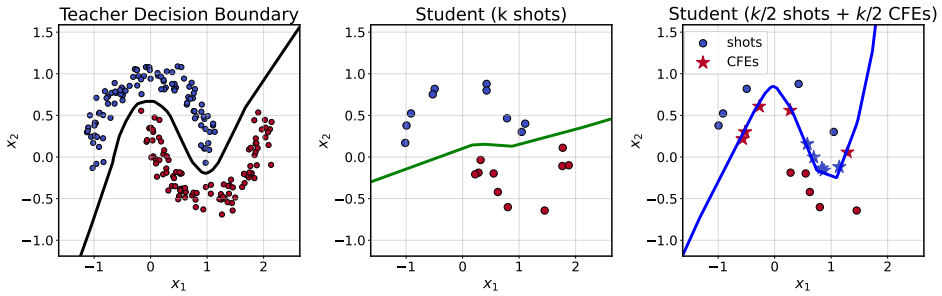


Figure 5.3: Decision boundaries for teacher and two student models trained on a synthetic 2D dataset under few-shot settings. The teacher *(a)* is trained on the full dataset and serves as the distillation target. First student *(b)* is distilled using 20 randomly sampled data points, and results in a poorly aligned decision boundary with the teacher. Second student *(c)* is also trained on 20 total samples, 10 original data points and their 10 CFEs. This student learns a decision boundary that aligns more closely with the teacher, as the KD loss encourages the student to match the teacher's soft predictions, guiding the CFEs to lie near the decision boundary.

The *student* network with a smaller architecture $[2 \to 16 \to 2]$. To simulate few-shot supervision, we randomly sample $k = 20$ original points (10 per class). For the original points, we compute their closest CFE (recall Definition 5.1), a minimally perturbed input that flips the teacher's predicted class. We follow a gradient-based method [25] to compute CFEs by perturbing each point in the direction of the teacher's logit margin until the predicted class flips. We consider two student models: one trained on the k few-shot samples alone, and another trained on $k/2$ few-shot samples and their CFEs. In both cases, we perform knowledge distillation by minimizing a combination of cross-entropy loss on the hard labels and KL-divergence between the student and teacher soft predictions. Figure 5.3 shows the decision boundaries of the teacher, the baseline student, and the CFE-infused student. CFEs cluster near the decision boundary, enriching the distillation data in high-uncertainty regions. The student trained with CFEs aligns more closely with the teacher, thus motivating the use of boundary-targeted examples for improved knowledge distillation.

5.4.2 Statistical Guarantees Motivating Our Approach

Here, we provide a theoretical motivation for the use of CFEs in few-shot knowledge distillation. We analyze a logistic regression setting using a measure from estimation theory called Fisher Information [217] (also see Definition 5.2) that captures the information contained by a random variable about a parameter to be estimated. We show that a dataset containing CFEs, which essentially lie much closer to the teacher's decision boundary, yields a Fisher Information Matrix with higher overall information content for parameter estimation. As a result, the student's expected estimation error is lower compared to training on standard samples.

Definition 5.2 (Fisher Information Matrix [217])**.** *Let $\mathcal{L}(\theta)$ be the log-likelihood of a parametric distribution $p(y, x; \theta)$, where θ is the parameter vector to be estimated. The* Fisher Information Matrix *(FIM) at parameter θ is defined as:*

$$\mathcal{I}(\theta) = \mathbb{E}_{\mathbf{x},y}\left[\nabla_\theta \log p(y, \mathbf{x}; \theta)\, \nabla_\theta \log p(y, \mathbf{x}; \theta)^\top\right].$$

Fisher Information measures the curvature of the log-likelihood: flatter regions (low curvature) imply high uncertainty in estimating θ, while sharper regions (high curvature) indicate that small changes in θ cause large changes in likelihood, enabling more precise parameter estimation. We consider a binary classification setting where the teacher and student are logistic regression models. The teacher parameterized by $\mathbf{w}_t$, defines the true data-generating distribution with predicted probabilities $p_t(y = 1|\mathbf{x}) = \sigma(\mathbf{w}_t^\top \mathbf{x})$ where $\sigma(\cdot)$ a softmax. Suppose, the student, with parameters $\mathbf{w}_s$, is obtained via maximum likelihood estimation (MLE) [217] using either a standard dataset $\mathcal{D}$ or CFE-infused dataset $\mathcal{D}_{\text{cf}}$. Since CFEs $\mathbf{x}_c$ lie close to the teacher's decision boundary, we have $\mathbf{w}_t^\top \mathbf{x}_c \approx 0$.

Theorem 5.1 (CFEs Improve Model Parameter Estimation)**.** *Let* $\mathbf{w}_s$ *and* $\mathbf{w}_s^{(\mathrm{cf})}$ *be the student parameters obtained via MLE on* $\mathcal{D}$ *(standard) and* $\mathcal{D}_{\mathrm{cf}}$ *(CFE-infused). Assuming the teacher's parameters* $\mathbf{w}_t$ *capture the true data-generating distribution, that CFEs lie near the decision boundary, and that the second moments* $\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top] \approx \mathbb{E}_{\mathbf{x}_c}[\mathbf{x}_c\mathbf{x}_c^\top]$*. Then estimation error satisfies:*

$$\mathbb{E}\left[\|\mathbf{w}_s^{(\mathrm{cf})} - \mathbf{w}_t\|^2\right] < \mathbb{E}\left[\|\mathbf{w}_s - \mathbf{w}_t\|^2\right].$$

Proof Sketch: The key step in our proof relies on showing that the Fisher Information is given by $\mathcal{I}(\mathbf{w}_t; \mathcal{D}) = \sum_i p_t(y=1|\mathbf{x}_i)(1-p_t(y=1|\mathbf{x}_i))\mathbf{x}_i\mathbf{x}_i^\top$. The scalar weight $p_t(y=1|\mathbf{x})(1-p_t(y=1|\mathbf{x}))$ is maximized when $p_t(y=1|\mathbf{x}) = 0.5$, i.e., $\mathbf{x}$ lies on the decision boundary. Standard samples in few-shot settings typically lie far from the boundary and contribute little to the FIM, whereas CFEs are constructed to lie near it and thus contribute significantly more. As a result, the FIM of the CFE-infused dataset $\mathcal{D}_{\mathrm{cf}}$ dominates that of the standard dataset $\mathcal{D}$ in Loewner order [218] (i.e., $\mathcal{I}(\mathbf{w}_t; \mathcal{D}_{cf}) \succ \mathcal{I}(\mathbf{w}_t; \mathcal{D})$). The CFE-infused dataset provides strictly more information for parameter estimation than the standard dataset, ultimately leading to the bound on expected estimation error (see proof in Appendix D.2).

While this result mathematically motivates the advantages of CFEs in few-shot distillation, it still assumes linear models and same student-teacher capacity (size). For more general non-linear settings, we provide a geometric perspective as discussed next.

5.4.3 Geometric Insight for Using CFEs for Distillation

Here, we examine the geometric effect of CFEs on student-teacher alignment in non-linear settings. Specifically, we show that when data points and their CFE pairs are included during distillation,

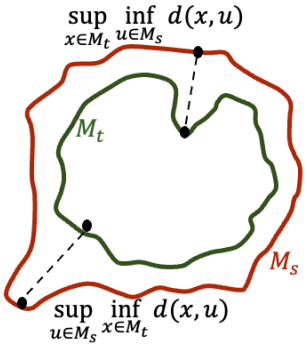


Figure 5.4: Hausdorff Distance measures how far two subsets of a metric space are from each other [3].

the student's decision boundary comes much closer to the teacher's boundary, as quantified by a formal measure called *Hausdorff distance* [3] between their respective decision surfaces. The Hausdorff distance (see Figure 5.4) captures the worst-case discrepancy between two sets (in our case, the decision boundaries of the teacher and student models) by quantifying how far any point on one boundary is from the closest point on the other.

Let $\mathcal{M}_t$=$\{\mathbf{x} \in \mathbb{R}^{n\times d} \mid f_t(\mathbf{x}) = 0.5\}$ and $\mathcal{M}_s$=$\{\mathbf{x} \in \mathbb{R}^{n\times d} \mid f_s(\mathbf{x}) = 0.5\}$ denote the decision boundaries of the teacher and student. Our goal is to examine how close is the student's decision boundary to the teacher's. To quantify this alignment, we define the Hausdorff distance as follows:

Definition 5.3 (Hausdorff Distance)**.** *Let* $\mathcal{M}_t, \mathcal{M}_s \subseteq \mathbb{R}^{n\times d}$ *be two non-empty subsets of a metric space. The* Hausdorff distance *is defined as:*

$$\mathrm{H}(\mathcal{M}_s, \mathcal{M}_t) = \max\Big\{ \sup_{\mathbf{x}\in\mathcal{M}_t} \inf_{\mathbf{u}\in\mathcal{M}_s} \|\mathbf{x}-\mathbf{u}\|_F,\ \sup_{\mathbf{u}\in\mathcal{M}_s} \inf_{\mathbf{x}\in\mathcal{M}_t} \|\mathbf{u}-\mathbf{x}\|_F \Big\}.$$

We observe that for each training sample $\mathbf{x}_i$ and its CFE $\mathbf{x}'_i$, the segment joining them cuts the teacher's decision boundary because they have different predictions. Essentially, there exists an intersection point $\mathbf{x}^\star_i$ on this segment such that $f_t(\mathbf{x}^\star_i) = 0.5$. Now, if the student is taught to match the

teacher at the sample $\mathbf{x}_i$ and its CFE $\mathbf{x}'_i$, the student would also have another intersection point on this segment. These two intersection points lying on the teacher and student decision boundaries will act as clamps, pulling the two boundaries close to each other, since their own gap gets smaller as $\mathbf{x}_i$ and its CFE $\mathbf{x}'_i$ comes closer. We assume boundaries are closed and distance is measured within a compact region (e.g., support of data).

Lemma 5.1 (Existence of Boundary Crossing for Counterfactual Pairs)**.** *Let $f_t : \mathbb{R}^{n\times d} \to [0,1]$ be a continuous function. For a datapoint and its counterfactual pair $(\mathbf{x}_i, \mathbf{x}'_i)$, there exists a point $\mathbf{x}^\star_i = \alpha\mathbf{x}_i + (1-\alpha)\mathbf{x}'_i$ for an $\alpha \in (0,1)$ (on the line joining $\mathbf{x}_i$ and $\mathbf{x}'_i$) such that: $f_t(\mathbf{x}^\star_i) = 0.5$.*

Theorem 5.2 (Teacher–Student Boundary Proximity)**.** *Let $f_t,\ f_s : \mathbb{R}^{n\times d} \to; [0,1]$ be the teacher and student model, with decision boundaries $\mathcal{M}_t = \{\mathbf{x} \mid f_t(\mathbf{x}) = 0.5\}$ and $\mathcal{M}_s = \{\mathbf{x} \mid f_s(\mathbf{x}) = 0.5\}$, respectively. Assume we observe a* CFE-infused dataset *$\mathcal{D}_{cf} = \big\{(\mathbf{x}_i, \mathbf{x}'_i)\big\}_{i=1}^{k}$ satisfying, for every pair $(\mathbf{x}_i, \mathbf{x}'_i)$: (A1) Minimal perturbation: $\|\mathbf{x}_i - \mathbf{x}'_i\|_F \le \alpha$ with $\alpha > 0$; (A2) Exact distillation: $f_s(\mathbf{x}_i) = f_t(\mathbf{x}_i)$ and $f_s(\mathbf{x}'_i) = f_t(\mathbf{x}'_i)$; and (A3) ε-spread along the teacher and student boundary, i.e., for each pair, there exist a teacher's crossing point $\mathbf{x}^\star_i = \alpha\mathbf{x}_i + (1-\alpha)\mathbf{x}'_i$ for $\alpha \in (0,1)$ such that $f_t(x^\star_i) = 0.5$ and for every $a \in \mathcal{M}_t$, there exists an i with $\|a - \mathbf{x}^\star_i\|_2 \le \varepsilon$. Then the Hausdorff distance between the decision boundaries obeys: $\mathrm{H}(\mathcal{M}_s, \mathcal{M}_t) \le \alpha + \varepsilon$.*

Consequently, tight (small α) and well-spread (small ε) CFE pairs guarantee that the student boundary remains inside an $(\alpha+\varepsilon)$-tube around the teacher boundary.

Interpretation of the assumptions and bound. Our theorem makes three intuitive assumptions. *(A1)* Minimal perturbation requires each input and its CFE pair $(\mathbf{x}, \mathbf{x}')$ to differ by at most α. CFEs are by definition the minimal changes that flips the teacher's prediction, so α is typically much smaller than the distance between arbitrary training points (note that we do no need CFEs to sit exactly on the

teacher's boundary, i.e., $f_t = 0.5$). It suffices that the perturbation is small and flips the label—capturing the practical way CFEs are produced. *(A2)* Exact distillation agreement assumes the student matches the teacher's outputs on the input and CFE pairs. This is reasonable, as these examples are directly used in training, and their logits are aligned through the distillation (KL) loss. *(A3)* ε-spread assumes the inputs are reasonably well spread. No region of the teacher's boundary is more than ε away from a crossing point. Under these assumptions, the Hausdorff gap between student and teacher boundaries is tightly bounded by $\alpha + \varepsilon$. This ensures the student's decision boundary stays within an $(\alpha + \varepsilon)$-tube around the teacher's, illustrating the geometric faithfulness we want in few-shot knowledge distillation. See proofs in Appendix D.4.

Proposed Algorithm (CoD). We propose CoD, a **Co**unterfactual Explanation-infused **D**istillation strategy for few-shot, task-aware distillation of LLMs. The first step is CFE generation. Existing methods primarily fall into optimization-based [25], search-based [219], and generative approaches [220]. These methods can be computationally expensive for LLMs, and frequently yield out-of-distribution or semantically implausible examples. To address this, we adopt a hybrid approach that combines the teacher model predictions with an LLM as an oracle for CFE generation. Specifically, given an input and its original label, we prompt an LLM (e.g., `GPT-4o` [221]) to generate a semantically similar sentence intended to flip the label with minimal changes to the input. We then check whether this generated example indeed flips the teacher model's prediction, ensuring its utility as a true CFE. Once validated, each CFE is paired with its original input $(\mathbf{x}, \mathbf{x}')$ and added to the training set. During distillation, we ensure that each input–CFE pair is included in the same mini-batch, enabling the student to jointly learn from both examples. The student is then trained using a combination of task loss, KL-based distillation loss, and optional layer-wise alignment. An overview of this process is described in Algorithm 4, with full implementation details and prompts provided in Appendix D.5.

5.5 Experiments

The goal of our experiments is to evaluate the effectiveness of integrating CFEs for knowledge distillation under few-shot learning settings. We investigate whether using limited real samples infused with their corresponding CFEs enable better distillation compared to only using real samples.

Algorithm 4 CoD: CFE-infused Distillation

Require: Teacher g_t, student g_s, data set $\mathcal{D}=\{(\mathbf{x}_i, y_i)\}_{i=1}^{k}$, CFGen, learning rate η, loss weights α (KD), β (LWD)

$\mathcal{D}_{\text{cf}} \leftarrow \emptyset$
for all $(\mathbf{x}, y) \in \mathcal{D}_k$ **do**
  $x' \leftarrow \text{CFGen}(\mathbf{x}, g_t)$
  $\mathcal{D}_{\text{cf}} \leftarrow \mathcal{D}_{\text{cf}} \cup \{(\mathbf{x}', 1-y)\}$
end for
$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_k \cup \mathcal{D}_{\text{cf}}$
for $e = 1$ **to** E **do**
  for all $(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}$ **do**
    $\mathcal{L}_{\text{hard}} \leftarrow \text{CE}(g_s(\mathbf{x}), y)$
    $\mathcal{L}_{\text{KD}} \leftarrow \text{KL}(g_t(\mathbf{x}) \,\|\, g_s(\mathbf{x}))$
    $\mathcal{L}_{\text{LWD}} \leftarrow \sum_{l \in \mathcal{I}} \|h_t^{(l)} - h_s^{(l)}\|_2^2$
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{hard}} + \alpha\, \mathcal{L}_{\text{KD}} + \beta\, \mathcal{L}_{\text{LWD}}$
    Update $\theta_s \leftarrow \theta_s - \eta\, \nabla_{\theta_s} \mathcal{L}$
  end for
end for
return distilled student g_s

Datasets. We evaluate CoD across six text classification benchmarks that span a range of domains. SST2 is a binary sentiment classification task derived from movie review snippets [222]. Sentiment140 consists of tweets labeled as positive or negative, reflecting user sentiment in short social media posts [223]. IMDB is a binary sentiment classification dataset containing full-length movie reviews [224]. CoLA (Corpus of Linguistic Acceptability) is a grammaticality judgment task that requires the model to identify whether a sentence is linguistically acceptable [225]. Amazon Polarity contains customer reviews labeled as positive or negative sentiment [226]. Yelp is another sentiment classification dataset based on user-generated restaurant reviews [227].

Model. We experiment with two prominent model families: DeBERTa-v3 [190] and Qwen2.5 [167]. For DeBERTa-v3, we use the "`base`" model (100M parameters) as the teacher and distill into two smaller "`small`" (44M) and "`xsmall`" (22M) variants as students. For Qwen2.5, we use Qwen2.5-1.5B as the teacher and distill into the smaller Qwen2.5-0.5B. Full training details are in Appendix D.5.

Baselines. We compare our method against three task-aware knowledge distillation baselines: (i) Standard knowledge distillation (KD) where the student learns from the teacher's soft predictions using KL divergence [228]; (ii) Layer-wise distillation (LWD), which extends KD by additionally aligning the student's intermediate hidden representations with those of the teacher using mean squared error [178]; and (iii) TED (Task-aware Layer-wise Distillation) which incorporates task-specific neural filters at each layer to selectively transfer task-relevant information from teacher to student [177]. All methods are evaluated under k-shot training settings, and student models are trained on identical few-shot splits to ensure a fair comparison (see details in Appendix D.5).

Setup. As in prior works on task-aware distillation [177], we first train a teacher model on the full training dataset to serve as a strong source of supervision. A student model is then initialized and distilled using only k datapoints, where $k \in \{8, 16, 32, 64, 128, 512\}$. We apply our strategy CoD to three standard distillation baselines: KD, LWD, and TED. For a fair comparison, CoD uses $k/2$ original samples and their $k/2$ corresponding CFE (a total of k shots) while the baseline methods are trained on k original samples. Performance is evaluated using accuracy on the test set for each dataset. All experimental results are averaged over five runs, with the mean and standard deviation reported. Results for the DeBERTa-v3-base teacher and DeBERTa-v3-small student are shown in Table 5.1, while results for the smaller DeBERTa-v3-xsmall student are in Appendix D.5. For experiments using the Qwen2.5-1.5B teacher and the Qwen2.5-0.5B student, see Table 5.3. We report the accuracy of teacher models trained on the full datasets in Table D.1 in Appendix D.5.

Table 5.1: Classification accuracy ($\pm$ std) across datasets with varying total training sizes k. For CoD, training data consists of $k/2$ standard and $k/2$ CFEs. Teacher model DeBERTa-v3-base and student model DeBERTa-v3-small.

		Total Samples (k)					
Dataset	**Method**	**8**	**16**	**32**	**64**	**128**	**512**
Amazon Polarity	KD	0.671 ±0.046	0.712 ±0.033	0.758 ±0.032	0.789 ±0.022	0.823 ±0.016	0.846 ±0.007
	+CoD	**0.758** ±0.027	**0.795** ±0.033	**0.819** ±0.035	**0.812** ±0.004	**0.837** ±0.014	**0.860** ±0.015
	LWD	0.676 ±0.090	0.738 ±0.033	0.777 ±0.009	0.809 ±0.015	**0.827** ±0.025	**0.842** ±0.019
	+CoD	**0.724** ±0.052	**0.779** ±0.056	**0.811** ±0.015	**0.828** ±0.015	0.816 ±0.020	0.841 ±0.013
CoLA	KD	0.693 ±0.062	0.707 ±0.029	0.721 ±0.012	0.747 ±0.005	0.758 ±0.009	0.771 ±0.003
	+CoD	**0.739** ±0.026	**0.755** ±0.017	**0.769** ±0.011	**0.769** ±0.016	**0.772** ±0.006	**0.791** ±0.004
	LWD	0.713 ±0.031	0.698 ±0.037	0.731 ±0.021	0.744 ±0.007	0.750 ±0.018	0.761 ±0.011
	+ CoD	**0.730** ±0.035	**0.744** ±0.031	**0.762** ±0.011	**0.752** ±0.009	**0.756** ±0.010	**0.784** ±0.003
IMDB	KD	0.714 ±0.047	0.817 ±0.028	0.875 ±0.027	0.896 ±0.008	**0.912** ±0.009	**0.917** ±0.006
	+ CoD	**0.835** ±0.078	**0.888** ±0.005	**0.890** ±0.011	**0.899** ±0.007	0.907 ±0.006	0.913 ±0.005
	LWD	0.760 ±0.046	0.836 ±0.045	0.875 ±0.024	0.889 ±0.013	0.905 ±0.008	**0.914** ±0.006
	+ CoD	**0.861** ±0.017	**0.886** ±0.011	**0.893** ±0.006	**0.898** ±0.005	0.905 ±0.010	0.913 ±0.010
SST2	KD	0.617 ±0.042	0.712 ±0.052	0.757 ±0.063	0.820 ±0.019	0.848 ±0.013	**0.899** ±0.007
	+ CoD	**0.719** ±0.063	**0.781** ±0.034	**0.821** ±0.013	**0.827** ±0.008	**0.853** ±0.015	0.892 ±0.018
	LWD	0.627 ±0.053	0.721 ±0.055	0.776 ±0.031	0.817 ±0.005	0.829 ±0.013	**0.892** ±0.012
	+ CoD	**0.694** ±0.079	**0.785** ±0.028	**0.832** ±0.011	**0.830** ±0.007	**0.835** ±0.012	0.880 ±0.020
Yelp	KD	0.714 ±0.058	0.817 ±0.031	0.855 ±0.021	**0.878** ±0.006	0.885 ±0.018	**0.916** ±0.007
	+ CoD	**0.740** ±0.094	**0.832** ±0.045	**0.860** ±0.018	0.874 ±0.006	**0.888** ±0.013	0.913 ±0.011
	LWD	0.733 ±0.070	0.832 ±0.026	0.857 ±0.011	0.868 ±0.006	0.881 ±0.017	**0.920** ±0.010
	+ CoD	**0.738** ±0.093	**0.865** ±0.010	**0.870** ±0.017	**0.871** ±0.019	**0.885** ±0.007	0.913 ±0.013
Sent140	KD	0.580 ±0.039	0.597 ±0.042	0.645 ±0.023	0.690 ±0.035	0.752 ±0.011	**0.802** ±0.006
	+ CoD	**0.629** ±0.036	**0.640** ±0.048	**0.731** ±0.022	**0.754** ±0.017	**0.778** ±0.007	0.784 ±0.019
	LWD	0.581 ±0.041	0.593 ±0.039	0.665 ±0.027	0.708 ±0.029	**0.751** ±0.009	**0.785** ±0.019
	+ CoD	**0.628** ±0.034	**0.652** ±0.038	**0.706** ±0.016	**0.741** ±0.014	0.729 ±0.063	0.760 ±0.023

Results and Analysis. Across all datasets, we observe that CoD significantly improves performance in the low-data regime, particularly when $k \leq 64$. For example, on Amazon Polarity with only 8 labeled examples, KD + CoD achieves 75.8% accuracy compared to 67.1% for standard KD (8.7 points improvement). Similarly, for IMDB at $k = 8$, LWD + CoD improves over standard LWD by more than 10 points (86.1% vs. 76.0%). As the number of labeled examples increases, the benefits of CFE augmentation diminish. At $k = 512$, the performance of standard and CoD becomes nearly identical in many cases. However, even in these larger settings, it is important to note that our method achieves comparable results while using only $k/2$ real samples and $k/2$ CFE, effectively halving the amount of labeled data required to reach similar performance. The effectiveness of CFEs varies by dataset. On CoLA, we observe consistent improvements across all k values for both KD and LWD, indicating that CFEs are well-aligned with the task's grammaticality decision boundary. In contrast, datasets like Sentiment140 show strong early gains. For datasets such as IMDB and SST2, CFE provides substantial improvements at low k, but underperforms slightly at $k = 512$, possibly due to redundancy. Among distillation methods, LWD generally performs on par with or slightly better than KD across most settings, with CoD offering similar relative improvements for both.

We also compare with TED [177] which has been found to work well with larger distillation datasets. We note that TED introduces additional complexity by requiring the training of task-specific filters prior to distillation. Interestingly, we find that TED does not consistently outperform classical methods like KD or LWD in the *few-shot* settings (see Table 5.2). Nonetheless, TED + CoD yields consistent gains over standard TED, demonstrating that our approach is broadly applicable. Our findings suggest that *simpler distillation approaches like KD or LWD are preferable when data is scarce*: they are easier to implement and, when combined with CoD, deliver much stronger performance gains without the overhead of filter training.

Table 5.2: Classification accuracy ($\pm$ std) with TED and TED + CoD across datasets and varying total training sizes k. For CoD, training data consists of $k/2$ standard and $k/2$ CFEs. Teacher model is DeBERTa-v3-base and student model is DeBERTa-v3-small.

		Total Samples (k)					
Dataset	**Method**	**8**	**16**	**32**	**64**	**128**	**512**
Amazon Polarity	TED	0.646 ±0.075	0.697 ±0.033	0.758 ±0.012	0.816 ±0.023	0.814 ±0.020	0.846 ±0.025
	+ CoD	**0.731** ±0.054	**0.754** ±0.056	**0.802** ±0.007	**0.818** ±0.013	**0.805** ±0.008	**0.848** ±0.010
CoLA	TED	**0.750** ±0.022	0.737 ±0.028	0.731 ±0.020	0.746 ±0.011	0.760 ±0.011	0.772 ±0.010
	+ CoD	0.748 ±0.028	**0.757** ±0.023	**0.767** ±0.021	**0.768** ±0.016	**0.780** ±0.007	**0.791** ±0.006
IMDB	TED	0.695 ±0.018	0.800 ±0.042	0.854 ±0.023	0.876 ±0.012	**0.908** ±0.009	**0.917** ±0.006
	+ CoD	**0.827** ±0.056	**0.879** ±0.003	**0.884** ±0.007	**0.887** ±0.010	0.895 ±0.010	0.916 ±0.005
SST2	TED	0.597 ±0.052	0.701 ±0.055	0.732 ±0.026	0.812 ±0.026	0.829 ±0.002	**0.904** ±0.006
	+ CoD	**0.658** ±0.087	**0.779** ±0.012	**0.813** ±0.017	**0.833** ±0.014	**0.836** ±0.030	0.879 ±0.011
Yelp	TED	0.699 ±0.048	0.815 ±0.014	0.846 ±0.020	0.869 ±0.012	**0.894** ±0.009	**0.914** ±0.012
	+ CoD	**0.742** ±0.095	**0.837** ±0.016	**0.868** ±0.018	**0.878** ±0.019	0.886 ±0.013	0.913 ±0.008

Ablations. *(1) On template designing and prompt choices*. We conducted an experiment varying four prompt templates for generating CFEs and observed that CoD is robust to prompt choices, showing low standard deviation across variants and consistently outperforming the KD baseline low shot cases (see Table D.4). This suggests CoD is not overly sensitive to prompt used. One possible direction could be to use automatic prompt generation methods are these are typically more compute-intensive. Given our already strong and stable performance using simple manually designed prompts, such complex techniques may not be necessary. *(2) Computational and Memory Requirements.* We assess the computational efficiency of CoD under varying few-shot budgets k using the `codecarbon` package [229] to track runtime and energy consumption (see Table D.5). *(3) Effect of soft-label supervision.* To study the role of soft-labels in our approach, we conduct an ablation removing or corrupting the teacher's soft labels. (see Table D.6). Removing the soft label term ($\alpha = 0$) leads to a substantial drop in performance across all shot levels. Although CoD still improves over KD in this setting, the

Table 5.3: Classification accuracy (± std) of Qwen2.5 on CoLA and Yelp datasets with varying training sizes k. For CoD training data consists of $k/2$ standard and $k/2$ CFEs. Teacher model is Qwen2.5-1.5B and student model is Qwen2.5-0.5B. Refer to Appendix D.5 for other datasets.

		Total Samples (k)					
Dataset	**Method**	**8**	**16**	**32**	**64**	**128**	**512**
CoLA	KD	0.681 ±0.012	0.676 ±0.023	0.668 ±0.042	0.654 ±0.032	0.676 ±0.020	0.732 ±0.014
	+ CoD	**0.683** ±0.016	**0.686** ±0.018	**0.697** ±0.015	**0.711** ±0.020	**0.736** ±0.017	**0.757** ±0.011
	LWD	0.681 ±0.012	0.657 ±0.031	0.678 ±0.018	0.650 ±0.039	0.636 ±0.029	0.712 ±0.014
	+ CoD	**0.682** ±0.018	**0.687** ±0.013	**0.704** ±0.010	**0.714** ±0.020	**0.719** ±0.022	**0.755** ±0.013
Yelp	KD	0.684 ±0.021	0.759 ±0.040	0.827 ±0.030	0.861 ±0.017	**0.887** ±0.012	**0.920** ±0.010
	+ CoD	**0.745** ±0.029	**0.779** ±0.048	**0.828** ±0.072	**0.886** ±0.007	0.883 ±0.010	0.916 ±0.008
	LWD	0.685 ±0.019	0.777 ±0.036	0.837 ±0.027	0.876 ±0.020	**0.898** ±0.008	**0.920** ±0.005
	+ CoD	**0.746** ±0.028	**0.778** ±0.035	**0.847** ±0.020	**0.876** ±0.014	0.883 ±0.010	0.909 ±0.009

gains are significantly reduced. This highlights that the soft label calibration from the teacher is a key contributor to the effectiveness of counterfactual explanation data. Additionally, when replacing soft labels with random values, performance degrades sharply, likely due to inconsistency with the hard labels, introducing conflicting supervision signals in the training objective.

5.6 Discussion

We introduced CoD, a novel approach for task-aware knowledge distillation in few-shot settings that leverages CFEs to enhance the data efficiency of knowledge distillation. Our results show that CoD consistently outperforms existing distillation approaches in low-data regimes. Importantly, we demonstrate that CoD can achieve improved performance over baselines while effectively using only half the number of original data, with the remainder consisting of generated CFEs. This finding has significant implications for reducing the cost of data collection in real-world scenarios where sourcing high-quality data is expensive or time-consuming [230]. Our approach offers an explanation-

driven perspective on distillation. By including CFE's, we implicitly highlight the key features most important to flipping a teacher's decision. This may help the student model reduce its reliance on spurious correlations, especially in few-shot settings. In effect, CFE's guide the student to attend to "why" a label changes, not just "what" the label is. This bridges explainability and compression, turning explanations into actionable data for knowledge distillation. As research increasingly focuses on getting more from less data [231, 232], future work could extend our approach to generative sequence-to-sequence models, enabling efficient distillation beyond classification. More broadly, our approach offers a path toward data-efficient LLM distillation from minimal data while reducing cost and maintaining performance.

Limitations. While the counterfactual-explanation-infused knowledge distillation (CoD) method demonstrates strong empirical performance, several limitations remain. First, generating counterfactual explanations (CFEs) introduces additional computational overhead compared to standard distillation approaches. Moreover, our current CFE generation strategy, which relies on prompting LLMs, does not guarantee that we would get the closest counterfactual (as defined in Definition 5.1), potentially limiting the precision of our distilled knowledge. Future work could explore alternate methods for generating closer and semantically valid CFEs. Additionally, as with knowledge distillation in general, CoD is inherently dependent on the quality of the teacher model. Any inaccuracies or biases present in the teacher's decision boundary may be inherited by the student. Addressing robustness to flawed teachers remains an important direction for future research.

Societal Impact. CoD offers several potential societal impacts, particularly in reducing the cost and effort associated with data collection [230]. By enabling the distillation of high-performance models with fewer data samples, this approach can significantly lower data collection costs, making machine learning more accessible in low-resource environments. This is especially valuable in industries where

data is often scarce and expensive to obtain [230, 233]. Moreover, by requiring fewer samples and targeting smaller student models, CoD contributes to more efficient model training, lower energy requirements, and scalable deployment. Our method leverages explanations as a tool for more effective model compression. In doing so, it bridges the gap between explainability and model compression.

Chapter 6: Explaining Fairness in Distributed Environments

6.1 Introduction

Federated learning (FL) is a framework where several parties (*clients*) collectively train machine learning models while retaining the confidentiality of their local data [234, 235]. With the growing use of FL in various high-stakes applications, such as finance, healthcare, recommendation systems, etc., it is crucial to ensure that these models do not discriminate against any group. [236]. While there are several methods to achieve group fairness in the centralized settings [237], these methods do not directly apply to a FL setting since each client only has access to their local dataset, and hence, is restricted to only performing local disparity mitigation.

Recent works [238–240] focus on finding models that are fair when evaluated on the entire dataset across all clients, a concept known as *global fairness*. E.g., several banks may decide to engage in FL to train a model that will determine loan qualifications without exchanging data among them. A globally fair model does not discriminate against any certain group when evaluated on the entire dataset across all the banks. On the other hand, *local fairness* considers the disparity of the model at each client (when evaluated on a client's local dataset). Local fairness is important as the models are ultimately deployed and used locally [241].

Global and local fairness can differ, particularly when the local demographics at a client differ from the global demographic across the entire dataset (data heterogeneity, e.g., a bank with primar-

ily younger customers). Prior studies have mainly focused on either global or local fairness, without always considering their interplay. Global and local fairness align when data is i.i.d. across clients [240, 241], but their interplay in other scenarios is not well-understood.

This work aims to provide a fundamental understanding of group fairness trade-offs in the FL setting. We first formalize the notions of Global and Local Disparity in FL using information theory. Next, we leverage a body of work within information theory called partial information decomposition (PID) to further identify three sources of disparity in FL that contribute to the Global and Local Disparity, namely, *Unique Disparity*, *Redundant Disparity*, and *Masked Disparity*. This information-theoretic decomposition is significant because it helps us derive fundamental limits on the trade-offs between Global and Local Disparity, particularly under data heterogeneity, and provides insights on when they agree and disagree. We introduce the *Accuracy and Global-Local Fairness Optimality Problem* (AGLFOP), a novel convex optimization that rigorously examines the trade-offs between accuracy and both global and local fairness. This framework establishes the theoretical limits of what any FL technique can achieve in terms of accuracy and fairness *given a dataset and client distribution.* This work provides a more nuanced understanding of the interplay between these two fairness notions that can better inform disparity mitigation techniques and their convergence and effectiveness in practice.

Our main contributions can be summarized as follows:

- **Partial information decomposition of Global and Local Disparity:** We first define Global Disparity as the mutual information $\mathrm{I}(Z;\hat{Y})$ where $\hat{Y}$ is a model's prediction and Z is the sensitive attribute (see Definition 6.2). Then, we show that Local Disparity can in fact be represented as the conditional mutual information $\mathrm{I}(Z;\hat{Y}|S)$ where S denotes the client (see Definition 6.3). We also demonstrate relationships between these information-theoretic quantifications and well-known fairness metrics

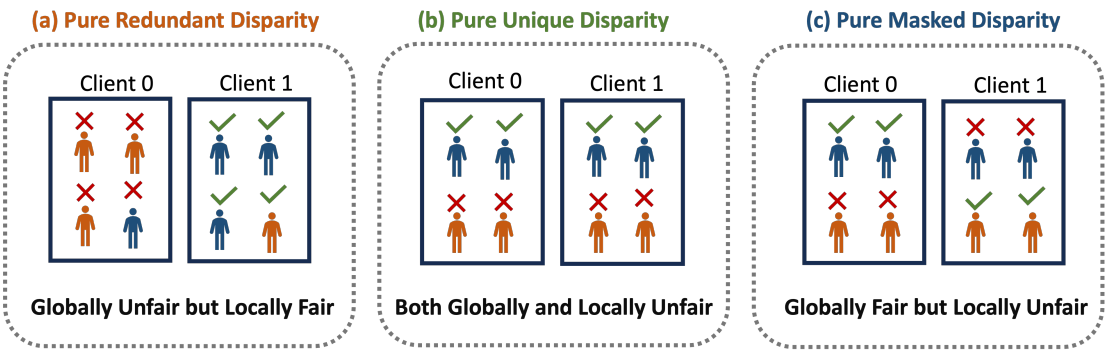


Figure 6.1: Illustrative scenarios of global and local disparities in FL. (*a*) Client 0 primarily serves younger adults, while Client 1 predominantly serves older adults. The model's prediction is entirely based on the client label. Within each client, both age groups are treated equally (locally fair), but globally, one group is disproportionately favored (globally unfair). (*b*) The model predictions are purely based on age, approving younger adults and rejecting older adults across both clients. Since age groups are uniformly distributed across clients, this results in both local and global unfairness. (*c*) The model approves younger adults from Client 0 and older adults from Client 1, while all others are rejected. At each client, the model exhibits a preference for a specific age group (locally unfair). However, globally, both groups receive equal approval rates (globally fair).

such as statistical parity (see Lemma 6.1). Using an information-theoretic quantification then enables us to further decompose the Global and Local Disparity into three non-negative components: *Unique*, *Redundant*, and *Masked Disparity*. We provide canonical examples to help understand these disparities in the context of FL (see Section 6.5.1). The significance of our information-theoretic decomposition lies in separating the regions of agreement and disagreement of Local and Global Disparity, demystifying their trade-offs.

- **Fundamental limits on trade-offs between local and global fairness:** We show the limitations of achieving global fairness using local disparity mitigation techniques due to *Redundant Disparity* (see Theorem 6.1) and the limitations of achieving local fairness even if global fairness is achieved due to Masked Disparity (see Theorem 6.2). We also identify the necessary and sufficient conditions under which one form of fairness (local or global) implies the other (see Theorem 6.3 and 6.4), as well as, discuss other conditions that are sufficient but not necessary.

- **A convex optimization framework for quantifying accuracy-fairness trade-offs:** We present the *Accuracy and Global-Local Fairness Optimality Problem* (AGLFOP) (see Definition 6.4), a novel convex optimization framework for systematically exploring the trade-offs between accuracy and both global and local fairness metrics. AGLFOP evaluates all potential joint distributions, thereby setting the theoretical boundaries for the best possible performance achievable for a given dataset and client distribution in FL.

- **Experimental demonstration:** We validate our theoretical findings using synthetic and real-world datasets. We study the trade-offs between accuracy and global-local fairness by examining the Pareto frontiers of the AGLFOP. We investigate the PID of disparities in the Adult dataset trained within a FL setting with multiple clients under various data heterogeneity scenarios.

6.2 Related Works

There are various perspectives to fairness in FL [242]. One definition is *client-fairness* [243], which aims to achieve equal performance across all client devices [244]. In this work, we are instead interested in group fairness, i.e., fairness with respect to demographic groups based on geographical location, etc. Methods for achieving group fairness in a centralized setting [4, 245–248] may not directly apply in a FL setting since each client only has access to their local dataset. Existing works on group fairness in FL generally aim to develop models that achieve *global fairness*, without much consideration for the *local fairness* at each client [240]. For instance, one approach to achieve global fairness in FL poses a constrained optimization problem to find the best model locally, while also ensuring that disparity at a client does not exceed a threshold and then aggregates those models [249–251]. Other techniques involve bi-level optimization that aims to find the optimal global model (minimum loss)

under the worst-case fairness violation [252–254], or re-weighting mechanisms [238, 239], both of which often require sharing additional parameters with a server. [241] argues for local fairness, as the model will be deployed at the local client level, and propose constrained multi-objective optimization. While accuracy-fairness tradeoffs have been examined in centralized settings [255–257, 257–263], such considerations, along with the relationship between local and global fairness, remain largely unexplored in FL. Our work addresses this gap by examining them through the lens of PID.

Information-theoretic measures have been used to quantify group fairness in the centralized setting in [22, 248, 264–274]. PID is also generating interest in other ML problems [275–281]. Here, instead of trying to minimize information-theoretic measures as a regularizer, our goal is to quantify the fundamental trade-offs between local and global fairness in FL and develop insights on their interplay to better understand what is information-theoretically possible using any technique.

6.3 Preliminaries

Notations. A client is represented as $S \in \{0, 1, \ldots, K-1\}$, where K is the total number of federating clients. A client $S{=}s$ has a dataset $\mathcal{D}_s = \{(x_i, y_i, z_i)\}_{i=1,\ldots n_s}$, where x_i denotes the input features, $y_i \in \{0, 1\}$ is the true label, and $z_i \in \{0, 1\}$ is the sensitive attribute (assume binary). The term n_s denotes the number of datapoints at client $S = s$. The collective dataset is given by $\mathcal{D} = \cup_{s=0}^{K-1} \mathcal{D}_s$. When denoting a random variable drawn from this dataset, we let X be the input features, Z be the sensitive attribute, and Y be the true label. We also let $\hat{Y}$ represent the predictions of a model $f_\theta(X)$ which is parameterized by θ.

Standard FL aims to minimize the empirical risk: $\min_\theta L(\theta) = \min_\theta \frac{1}{K} \sum_{s=0}^{K-1} \alpha_s L_s(\theta)$, where $L_s(\theta) = \frac{1}{n_s} \sum_{(x,y)\in\mathcal{D}_s} l(f_\theta(x), y)$ is the local objective (or loss) at client s, α_s is an importance co-

efficient (often equal across clients), and $l(\cdot, \cdot)$ denotes a predefined loss function. To minimize the objective $L(\theta)$, a decentralized approach is employed. Each client $S = s$ trains on their private dataset $\mathcal{D}_s$ and provides their trained local model to a centralized server. The server aggregates the parameters of the local models to create a global model $f_\theta(x)$ [282]. E.g., the *FedAvg* algorithm [283] is a popular approach that aggregates the parameters of local models by taking their average, which is then used to update the global model. This process is repeated until the global model achieves a satisfactory *performance* level.

6.4 Background on Partial Information Decomposition

PID decomposes the mutual information $\mathrm{I}(Z; A, B)$ about a random variable Z contained in the tuple (A, B) into four non-negative terms:

$$\mathrm{I}(Z; A, B) = \mathrm{Uni}(Z{:}A|B) + \mathrm{Uni}(Z{:}B|A) + \mathrm{Red}(Z{:}A, B) + \mathrm{Syn}(Z{:}A, B) \tag{6.1}$$

Here, $\mathrm{Uni}(Z{:}A|B)$ denotes the unique information about Z that is present only in A and not in B. E.g., *shopping preferences* (A) may provide unique information about *age* (Z) that is not present in *address* (B). $\mathrm{Red}(Z{:}A, B)$ denotes the redundant information about Z that is present in both A and B. E.g., *zipcode* (A) and *county* (B) may provide redundant information about *income level* or *age*. $\mathrm{Syn}(Z{:}A, B)$ denotes the synergistic information not present in either A or B individually, but present jointly in (A, B), e.g., each individual digit of the *zipcode* may not have information about *income level* or *age* but together they provide significant information about the feature.

Numerical Example. Let $Z = (Z_1, Z_2, Z_3)$ with each $Z_i \sim$ i.i.d. Bern(½). Let $A = (Z_1, Z_2, Z_3 \oplus N)$, $B = (Z_2, N)$, $N \sim$ Bern(½) is independent of Z. Here, $\mathrm{I}(Z; A, B) = 3$ bits. The unique

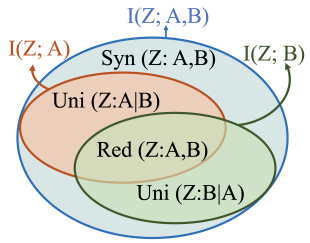


Figure 6.2: Venn diagram showing PID of mutual information $\mathrm{I}(Z; A, B)$.

information about Z that is contained only in A and not in B is effectively in Z_1, and is given by $\mathrm{Uni}(Z{:}A|B){=}\mathrm{I}(Z; Z_1) = 1$ bit. The redundant information about Z that is contained in both A and B is effectively in Z_2 and is given by $\mathrm{Red}(Z{:}A, B) = \mathrm{I}(Z; Z_2) = 1$ bit. Lastly, the synergistic information about Z that is not contained in either A or B alone, but is contained in both of them together is effectively in the tuple $(Z_3 \oplus N, N)$, and is given by $\mathrm{Syn}(Z{:}A, B) = \mathrm{I}(Z; (Z_3 \oplus N, N)) = 1$ bit. This accounts for the 3 bits in $\mathrm{I}(Z; A, B)$.

Definition 6.1 (Unique Information [284])**.** *Let Δ be the set of joint distributions on (Z, A, B) and Δ_p be the set of joint distributions with the same marginals on (Z, A) and (Z, B) as the true distribution, i.e., $\Delta_p = \{Q{\in}\Delta : \mathrm{Pr}_Q(Z{=}z, A{=}a){=}\mathrm{Pr}(Z{=}z, A{=}a)$ and $\mathrm{Pr}_Q(Z{=}z, B{=}b) = \mathrm{Pr}(Z{=}z, B{=}b)\}$. Then:*

$$\mathrm{Uni}(Z{:}A|B) = \min_{Q \in \Delta_p} \mathrm{I}_Q(Z; A|B)$$

where $\mathrm{I}_Q(Z; A|B)$ is the conditional mutual information when (Z, A, B) have joint distribution Q and $\mathrm{Pr}_Q(\cdot)$ denotes the probability under Q.

Defining any one of the PID terms suffices to obtain the others. $\mathrm{Red}(Z{:}A, B)$ is the sub-volume between $\mathrm{I}(Z; A)$ and $\mathrm{I}(Z; B)$ (see Fig. 6.2). Hence, $\mathrm{Red}(Z{:}A, B) = \mathrm{I}(Z; A) - \mathrm{Uni}(Z{:}A|B)$ and $\mathrm{Syn}(Z{:}A, B) = \mathrm{I}(Z; A, B) - \mathrm{Uni}(Z{:}A|B) - \mathrm{Uni}(Z{:}B|A) - \mathrm{Red}(Z{:}A, B)$ (from Eqn. (6.1)).

6.5 Main Contributions

We first formalize the notions of Global and Local Disparity in FL using information theory.

Definition 6.2 (Global Disparity)**.** *The Global Disparity of a model f_θ with respect to Z is defined as $\mathrm{I}(Z;\hat{Y})$, the mutual information between Z and $\hat{Y}$ (where $\hat{Y} = f_\theta(X)$).*

This is related to a widely-used group fairness notion called statistical parity. Existing works define the Global Statistical Parity as: $\Pr(\hat{Y} = 1|Z = 1) = \Pr(\hat{Y} = 1|Z = 0)$ [245]. Global Statistical Parity is satisfied when Z is independent of $\hat{Y}$, which is equivalent to zero mutual information $\mathrm{I}(Z;\hat{Y}) = 0$. To further justify our choice of $\mathrm{I}(Z;\hat{Y})$ as a measure of Global Disparity, we provide a relationship between the absolute statistical parity gap and mutual information when they are non-zero in Lemma 6.1 (Proof in Appendix E.2).

Lemma 6.1 (Relationship between Global Statistical Parity Gap and $\mathrm{I}(Z;\hat{Y})$)**.** *Let $\Pr(Z{=}0) = \alpha$. The gap $SP_{global} = |\Pr(\hat{Y} = 1|Z = 1) - \Pr(\hat{Y} = 1|Z = 0)|$ is bounded by $\frac{\sqrt{0.5\,\mathrm{I}(Z;\hat{Y})}}{2\alpha(1-\alpha)}$.*

A critical observation that we make in this work is that: *local unfairness can be quantified as $\mathrm{I}(Z;\hat{Y}|S)$, the conditional mutual information between Z and $\hat{Y}$ conditioned on S. This is motivated from [240] which defines Local Statistical Parity at a client s as: $\Pr(\hat{Y}{=}1|Z{=}1, S{=}s) = \Pr(\hat{Y}{=}1|Z{=}0, S{=}s)$.*

Definition 6.3 (Local Disparity)**.** *The Local Disparity is the conditional mutual information $\mathrm{I}(Z;\hat{Y}|S)$.*

Lemma 6.2. *$\mathrm{I}(Z;\hat{Y}|S){=}0$ if and only if $\Pr(\hat{Y}{=}1|Z{=}1, S{=}s){=}\Pr(\hat{Y}{=}1|Z{=}0, S{=}s)$ at all clients s.*

The proof (see Appendix E.2) uses the fact that $\mathrm{I}(Z;\hat{Y}|S) = \sum_{s=0}^{K-1} \Pr(S{=}s)\mathrm{I}(Z;\hat{Y}|S = s)$ where $\mathrm{I}(Z;\hat{Y}|S = s)$ is the Local Disparity at client s, and $\Pr(S{=}s) = n_s/n$, the proportion of data

points at client s. Similar to Lemma 6.1, we can also get a relationship between SP_s and $\mathrm{I}(Z;\hat{Y}|S=s)$ when they are non-zero (see Corollary E.1 in Appendix E.2). We can also define other fairness metrics similarly. For instance, *Global Equalized Odds* can be formulated in terms of the conditional mutual information, denoted as $\mathrm{I}(Z;\hat{Y}|Y)$ and *Local Equalized Odds* as $\mathrm{I}(Z;\hat{Y}|Y,S)$.

6.5.1 Partial Information Decomposition of Global and Local Disparity

We provide a decomposition of Global and Local Disparity into three sources of unfairness: Unique, Redundant, and Masked Disparity, and provide examples to illustrate and better understand these disparities in the context of FL.

The Global and Local Disparity in FL can be decomposed into non-negative terms:

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Red}(Z{:}\hat{Y},S). \tag{6.2}$$

$$\mathrm{I}(Z;\hat{Y}|S) = \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Syn}(Z{:}\hat{Y},S). \tag{6.3}$$

We refer to Fig. 6.3 for an illustration of this result. Eq. (6.2) follows from the relationship between different PID terms while Eq. (6.3) requires the chain rule of mutual information [285]. For completeness, we show the non-negativity of PID terms in Appendix E.3.

The term $\mathrm{Uni}(Z{:}\hat{Y}|S)$ quantifies the unique information the sensitive attribute Z provides about the model prediction $\hat{Y}$ that is not provided by client label S. We refer to this as the ***Unique Disparity***. The Unique Disparity contributes to both Local and Global Disparity, highlighting the region where they agree. The ***Redundant Disparity***, $\mathrm{Red}(Z{:}\hat{Y},S)$, quantifies the information about sensitive attribute Z that is common between prediction $\hat{Y}$ and client S. The Unique and Redundant Disparities together make up the Global Disparity. $\mathrm{Syn}(Z{:}\hat{Y},S)$ represents the synergistic information about

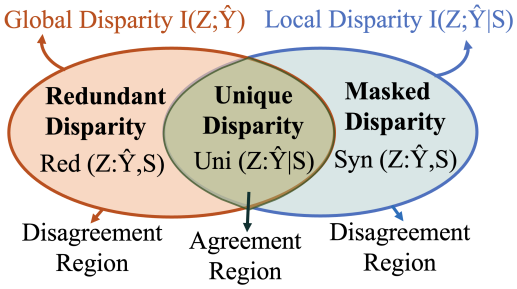


Figure 6.3: Venn diagram of PID for Global & Local Disp. with agreement and disagreement regions.

sensitive attribute Z that is *not* present in either $\hat{Y}$ or S individually, but is present jointly in $(\hat{Y}, S)$. We refer to this as the ***Masked Disparity***, as it is only observed when $\hat{Y}$ and S are considered together. Redundant and Masked Disparities cause disagreement between global and local fairness.

Canonical Examples. We now examine a loan approval scenario featuring binary sensitive attributes across two clients, i.e., $\hat{Y}, Z, S \in \{0, 1\}$. Here, $\mathrm{I}(Z; \hat{Y}, S) = H(Z) - H(Z|\hat{Y}, S) \leq H(Z) = 1$, i.e., the maximum disparity is 1 bit.

Example 6.1 (Pure Uniqueness)**.** *Let $\hat{Y} = Z$ and $Z \perp\!\!\!\perp S$. The younger adults ($Z = 1$) and older adults ($Z = 0$) are identically distributed across the clients. Suppose, the model only approves younger adults but rejects older adults for a loan across both clients. This model is both locally and globally unfair, $\mathrm{I}(Z; \hat{Y}) = \mathrm{I}(Z; \hat{Y}|S) = 1$. This is a case of purely Unique Disparity since all the information about age is derived exclusively from the model predictions; the client S has no correlation with age Z. Both Global and Local Disparities are in agreement. Here, $\mathrm{Uni}(Z{:}\hat{Y}|S) = 1$, $\mathrm{Red}(Z{:}\hat{Y}, S) = 0$, and $\mathrm{Syn}(Z{:}\hat{Y}, S) = 0$.*

Example 6.2 (Pure Redundancy)**.** *The client $S = 0$ has 90% older adults, while client $S = 1$ has 90% younger adults. So, there is a correlation between the clients and age. Suppose, the model approves everyone from client $S = 1$ while rejecting everyone in $S = 0$ (i.e., $\hat{Y} = S$). Such a model*

is locally fair because younger and older adults are treated equally within a particular client, and $\mathrm{I}(Z;\hat{Y}|S) = 0$. *However, the model is globally unfair since* $\mathrm{I}(Z;\hat{Y}) = 0.53$. *This is a case with pure Redundant Disparity since information about* Z *is derived from both* $\hat{Y}$ *and* S. *Global and Local Disparities are in disagreement. Here,* $\mathrm{Uni}(Z{:}\hat{Y}|S) = 0$, $\mathrm{Red}(Z{:}\hat{Y}, S) = 0.53$, *and* $\mathrm{Syn}(Z{:}\hat{Y}, S) = 0$.

In general, pure Redundant Disparity is observed when Z and $\hat{Y}$ are correlated and $Z - S - \hat{Y}$ form a *Markov chain*, i.e., $\hat{Y} = S$ and $S = g(Z)$ for some function g.

Example 6.3 (Pure Synergy)**.** *Let* $\hat{Y} = Z \oplus S$ *and* $Z \perp\!\!\!\perp S$. *The model approves younger adults (*$Z = 1$*) from client* $S = 0$ *and older adults (*$Z = 0$*) from client* $S = 1$, *while others are rejected. Such a model is locally unfair, as it singularly prefers one age group within each client with* $\mathrm{I}(Z;\hat{Y}|S) = 1$. *However, it is globally fair since it maintains an equal approval rate for both younger and older adults with* $\mathrm{I}(Z;\hat{Y}) = 0$. *This is a case with pure Masked Disparity as information about* Z *that is not observable in either* $\hat{Y}$ *or* S *individually is present jointly. Here,* $\mathrm{Uni}(Z{:}\hat{Y}|S) = 0$, $\mathrm{Red}(Z{:}\hat{Y}, S) = 0$, *and* $\mathrm{Syn}(Z{:}\hat{Y}, S) = 1$.

Merits of PID. These canonical examples demonstrate scenarios with pure uniqueness, redundancy, and synergy. In practice, there is usually a mixture of all of these disparities. i.e., non-zero Unique, Redundant, and Masked Disparities. In these scenarios, PID serves as a powerful tool that can disentangle the regions of agreement and disagreement between Local and Global Disparity, particularly when data is distributed non-identically across clients (also see experiments in Section 6.6). In contrast, traditional fairness metrics lack the granularity to capture these nuanced interactions, making PID an essential asset for a more comprehensive understanding and mitigation of disparities. Using PID, we can uncover the fundamental information-theoretic limits and trade-offs between Global and Local Disparities, which we will examine in greater depth next.

6.5.2 Fundamental Limits on Tradeoffs Between Local and Global Disparity

We examine the use of local fairness to achieve global fairness, or scenarios where a model is trained to achieve local fairness and subsequently deployed at the global level. Since clients have direct access only to their data, implementing local disparity mitigation techniques at the individual client level is both practical and convenient. Studies such as [241] argue that local fairness is important as models are deployed at the local client level. However, this raises a critical question about the impact on global fairness. In Theorem 6.1, we formally demonstrate that even if local clients are able to use some optimal local mitigation methods and model aggregation techniques to achieve local fairness, the Global Disparity may still be greater than zero.

Theorem 6.1 (Impossibility of Using Local Fairness to Attain Global Fairness)**.** *As long as Redundant Disparity* $\mathrm{Red}(Z{:}\hat{Y}, S) > 0$*, the Global Disparity* $\mathrm{I}(Z; \hat{Y}) > 0$ *even if Local Disparity goes to* 0*.*

In order to achieve local fairness, Unique and Masked Disparities must be reduced to zero. The proof leverages Proposition 6.5.1, particularly relying on non-negativity of Unique and Redundant Disparities (see Appendix E.4). Recall, Example 6.2 (Pure Redundancy), where the Local Disparity was zero, but the Global Disparity was 0.53 as a result of the Redundant Disparity. This is not uncommon in real-world scenarios. Sensitive attributes like age may be correlated with location. For instance, one hospital may mainly cater to younger patients, whereas another could predominantly serve older patients. A model may be trained to achieve local fairness but would fail to be globally fair due to a non-zero Redundant Disparity, highlighting the region of disagreement (see Fig. 6.3).

We now consider the scenario where a model is able to achieve global fairness and is subsequently deployed at the local client level.

Theorem 6.2 (Global Fairness Does Not Imply Local Fairness)**.** *As long as Masked Disparity* $\mathrm{Syn}(Z{:}\hat{Y}, S) > 0$*, local fairness will not be attained even if global fairness is attained.*

To achieve global fairness, the Unique and Redundant Disparities must reduce to zero. Recall Example 6.3 (Pure Synergy), where the model accepts younger adults from client $S = 0$ and older adults from client $S = 1$, while rejecting all others. While this model is globally fair, it is not locally fair. This demonstrates that while it is possible to train a model to achieve global fairness, it may still exhibit disparity when deployed at the local level due to the canceling of disparities between clients. This effect is captured by the Masked Disparity. We now discuss the necessary and sufficient conditions to achieve global fairness using local fairness.

Theorem 6.3 (Necessary and Sufficient Condition to Achieve Global Fairness Using Local Fairness)**.** *If Local Disparity* $\mathrm{I}(Z; \hat{Y}|S)$ *goes to zero, then Global Disparity* $\mathrm{I}(Z; \hat{Y})$ *also goes to zero, if and only if the Redundant Disparity* $\mathrm{Red}(Z{:}\hat{Y}, S){=}0$*. A sufficient condition for* $\mathrm{Red}(Z{:}\hat{Y}, S){=}0$ *is* $Z \perp\!\!\!\perp S$*.*

Theorem 6.3 suggests that if the sensitive attribute is uniformly distributed across clients the Redundant Disparity will reduce to zero (see proof in Appendix E.4). Hence, when the Local Disparity goes to zero, the Global Disparity will also decrease to zero. However, in practice, this proportion is fixed since the dataset at each client cannot be changed, i.e., $\mathrm{I}(Z; S)$ is fixed. Therefore, we examine another more controllable condition to eliminate Redundant Disparity even when $\mathrm{I}(Z; S) > 0$.

One might think that a potential solution to have $\mathrm{Red}(Z{:}\hat{Y}, S) = 0$ is to enforce independence between $\hat{Y}$ and S, i.e., the model should make predictions at the same rate across all clients. However, interestingly, the PID literature demonstrates counterexamples [286] where this does not hold. We show that an additional condition of $\mathrm{Syn}(Z{:}\hat{Y}, S) = 0$ is required.

A sufficient condition for $\mathrm{Red}(Z{:}\hat{Y}, S) = 0$ is $\mathrm{Syn}(Z{:}\hat{Y}, S) = 0$ and $\hat{Y} \perp\!\!\!\perp S$.

Remark 6.1. *It is worth noting that the independence between $\hat{Y}$ and S can be approximately achieved if the true Y and S are independent, as $\hat{Y}$ is an estimation of Y. The mutual information $\mathrm{I}(Y;S)$ can provide insights into the anticipated value of $\mathrm{I}(\hat{Y};S)$, as FL typically aims to also achieve a reasonable level of accuracy. However, it is often the case that $\mathrm{I}(\hat{Y};S)$ is fixed due to the fixed nature of datasets at each client. It may even be possible to enforce $\hat{Y} \perp\!\!\!\perp S$ at the cost of accuracy.*

Lastly, we examine conditions to attain local fairness through global fairness.

Theorem 6.4. *Local disparity will always be less than Global Disparity if and only if Masked Disparity $\mathrm{Syn}(Z{:}\hat{Y}, S) = 0$. A sufficient condition is when $Z - \hat{Y} - S$ form a Markov chain.*

Remark 6.2 (Extension to Personalized Federated Learning Setting)**.** *Interestingly, our results extend to the personalized FL setting, where client s can tailor the final global model $\hat{Y} = f(X)$ into a personalized version to improve local performance, $\hat{Y} = f_s(X)$. In this case, we can define the global model as a random variable $\hat{Y} = g(X, S)$ and all of the propositions would hold.*

6.5.3 An Optimization Framework for Exploring the Accuracy Fairness Trade-off

We investigate the inherent trade-off between model accuracy and fairness in the FL context. We formulate the *Accuracy and Global-Local Fairness Optimality Problem (AGLFOP)*, an optimization to delineate the theoretical boundaries of accuracy and fairness trade-offs, capturing the optimal performance any model or FL technique can achieve for a specified dataset and client distribution.

Let Δ be the set of all joint distributions defined for $(Z, S, Y, \hat{Y})$. Let Δ_p be a set of all joint distributions $Q \in \Delta$ that maintain fixed marginals on (Z, S, Y) as determined by a given dataset and client distribution, i.e., $\Delta_p = \{Q \in \Delta : \Pr_Q(Z{=}z, S{=}s, Y{=}y) = \Pr(Z{=}z, S{=}s, Y{=}y), \forall z, s, y\}$.

Definition 6.4 (Accuracy and Global-Local Fairness Optimality Problem (AGLFOP))**.** *Let* $\mathrm{I}_Q(Z;\hat{Y})$ *and* $\mathrm{I}_Q(Z;\hat{Y}|S)$ *be Global and Local Disparity under distribution* Q*. Then, the AGLFOP for a specific dataset and client distribution is an optimization of the form:*

$$\underset{Q\in\Delta_p}{\arg\min} \quad err(Q) \quad \textit{subject to} \quad \mathrm{I}_Q(Z;\hat{Y}) \leq \epsilon_g, \quad \mathrm{I}_Q(Z;\hat{Y}|S) \leq \epsilon_l, \tag{6.4}$$

where $err(Q)=\sum_{z,s,y,\hat{y}} \Pr_Q(Z{=}z, S{=}s, Y{=}y, \hat{Y}{=}\hat{y})\mathbb{I}(y \neq \hat{y})$*, the classification error under distribution* Q *(*$\mathbb{I}(\cdot)$ *denotes the indicator function).* $err(Q) \in [0,1]$ *quantifies the proportion of incorrect predictions, calculated as the summation of the probabilities of misclassifying the true labels. The complement of the classification error,* $1 - err(Q)$*, quantifies the accuracy.*

Theorem 6.5. *The AGLFOP is a convex optimization problem.*

The AGLFOP is a convex optimization problem (see proof in Appendix E.5) that evaluates all potential joint distributions within the set Δ_p which includes the specific dataset and how they are distributed across clients. The true distribution of this given dataset across clients is $\Pr(Z = z, S = s, Y = y)$. This makes it an appropriate framework for investigating the accuracy-fairness trade-off. The Pareto front of this optimization problem facilitates a detailed study of the trade-offs, showcasing the maximum accuracy that can be attained for a given global and local fairness relaxation (ϵ_g, ϵ_l).

The set Δ_p can be further restricted for specialized applications, e.g., to constrain to all derived classifiers from an optimal classifier [246]. We can restrict our optimization space Δ_p to lie within the convex hull derived by the False Positive Rate (FPR) and True Positive Rate (TPR) of an initially trained classifier. This would characterize the accuracy-fairness tradeoff for all derived classifiers from the original trained classifier. The convex hall characterizes the distributions that can be achieved with

any derived classifier. The constraints of AGLFOP can also be expressed using PID terms, offering intriguing insights that we will examine in the following section. The AGLFOP can be computed in any FL environment. Specifically, their computation necessitates the characterization of the joint distribution $\Pr(Z{=}z, S{=}s, Y{=}y) = \Pr(S{=}s)\Pr(Z{=}z|S{=}s)\Pr(Y{=}y|Z{=}z, S{=}s)$, which can be readily acquired by aggregating pertinent statistics across all participating clients.

Remark 6.3 (Broader Potential)**.** *The AGLFOP currently focuses on independence between Z and $\hat{Y}$, but can be adapted to explore other fairness notions. It can also be used to study optimality in situations where different clients have varying fairness requirements, e.g., adhering to statistical parity globally while upholding equalized odds at the local level. Moreover, variants of this optimization problem can be developed to penalize only the worst-case client fairness scenarios.*

6.6 Experiments

In this section, we provide experimental evaluations on synthetic and real-world datasets to validate our theoretical findings.[1] We investigate the PID of Global and Local Disparity under various conditions. We also examine the trade-offs between these fairness metrics and model accuracy.

Data and Client Distribution. We consider the following: *(1) Synthetic dataset:* A 2-D feature vector $X{=}(X_0, X_1)$ has a distribution, i.e., $X|_{Y=1}{\sim}\mathcal{N}((2,2), \left[\begin{smallmatrix}5 & 1\\ 1 & 5\end{smallmatrix}\right])$, $X|_{Y=0}{\sim}\mathcal{N}((-2,-2), \left[\begin{smallmatrix}10 & 1\\ 1 & 3\end{smallmatrix}\right])$. Assume binary sensitive attribute $Z{=}1$ if $X_0{>}0$ and 0 otherwise to encode a dependence; and *(2) Adult dataset* [37] with a binary sensitive attribute. We consider three cases for partitioning the datasets across clients: (*Scenario* 1) sensitive-attribute independently distributed across clients, i.e., $Z \perp\!\!\!\perp S$, (*Scenario* 2) high sensitive-attribute heterogeneity across clients, i.e., $Z = S$ with probability α, and

[1] Implementation is available at `https://github.com/FaisalHamman/FairFL-PID`

(*Scenario* 3) high sensitive-attribute synergy level across clients, i.e., $Y = Z \oplus S$. Further details are described in Appendix E.6.

Experiment A: Accuracy-Global-Local-Fairness Trade-off Pareto Front. To study the trade-offs between model accuracy and different fairness constraints, we plot the Pareto frontiers for the AGLFOP. We solve for maximum accuracy $(1 - err)$ while varying global and local fairness relaxations (ϵ_g, ϵ_l). We present results for synthetic and Adult datasets as well as PID terms for various data splitting scenarios across clients.[2] The three-way trade-off among accuracy, global, and local fairness can be visualized as a contour plot (see Fig. 6.4).

Interestingly, PID allows us to quantify the agreement and disagreement between local and global fairness. In scenarios characterized by Unique Disparity, local and global fairness agree, and accuracy trade-offs are balanced between them. In cases characterized by Redundant Disparity, the trade-off is primarily between accuracy and global fairness (the accuracy changes along the horizontal axis (ϵ_l) seemingly nonexistent given (ϵ_g)). In contrast, scenarios with Masked Disparity exhibit a trade-off that is primarily between accuracy and local fairness (the trade-off is across the vertical axis).

Experiment B: Demonstrating Disparities in Federated Learning Settings. In this experiment, we investigate the PID of disparities on the Adult dataset trained within a FL framework. We employ the *FedAvg* algorithm [283] for training and analyze:

- *PID Across Various Splitting Scenarios.* We partition the dataset among clients based on Scenarios 1-3, utilize FedAvg for model training in each case, and examine the PID of both Local and Global Disparities (see Fig. 6.5). For each scenario, we also evaluate the effects of using a local disparity mitigation technique. This is achieved by incorporating a statistical parity regularizer at each client. The results and implementation details are presented in Table E.1 in Appendix E.6.

[2] We use Python *dit* package [287] for PID computation and *cvxpy* for convex solvers.

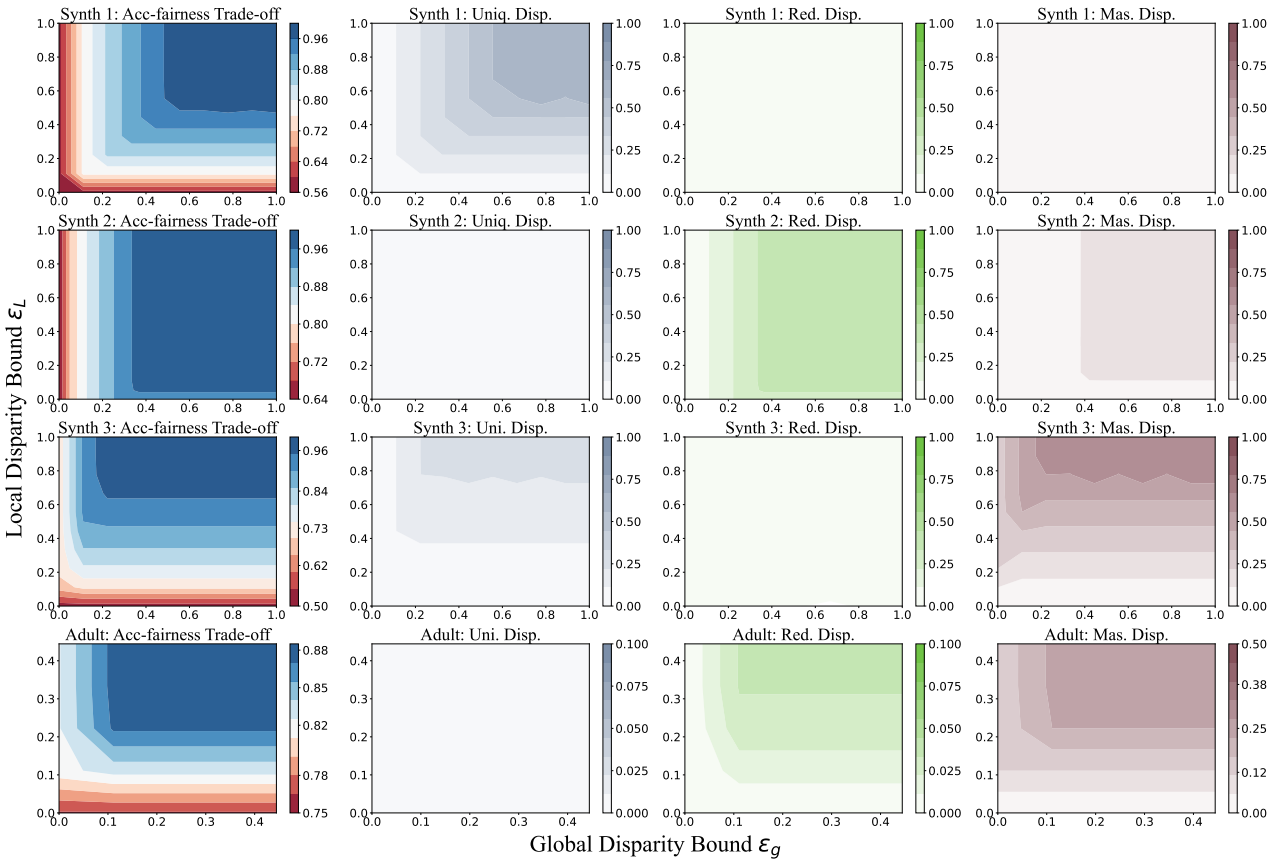


Figure 6.4: AGLFOP Pareto Frontiers for Synthetic and Adult Datasets with PID. (*first column*) shows maximum accuracy $(1 - err)$ that can be achieved on a dataset and client distribution for a given global and local fairness relaxation (ϵ_g, ϵ_l). Synthetic data in scenario 1 *(first row)* is characterized by Unique Disparity. Local and global fairness agree, and accuracy trade-offs are balanced between them. Synthetic data in scenario 2 with $\alpha = 0.9$ *(second row)* is dominated by Redundant Disparity with trade-offs mainly between global fairness and accuracy (an accurate model could have zero Local Disparity but be globally unfair). Synthetic data in Scenario 3 *(third row)* is characterized by Masked Disparity with trade-offs mainly between local fairness and accuracy (an accurate model could have zero Global Disparity but be locally unfair). Adult data with heterogeneous split *(fourth row; details in Appendix E.6)*, displaying predominantly Masked Disparity but notable presence of Redundant Disparity, capturing more complex relationships and trade-offs.

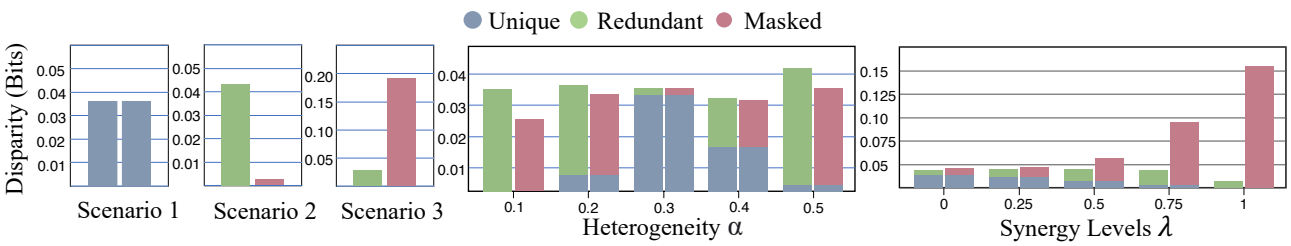


Figure 6.5: *(left)* Plot demonstrating scenarios with Unique, Redundant, and Masked Disparities for the Adult dataset (model trained using *FedAvg*). Unique Disparity when sensitive attributes are equally distributed across clients. Redundant Disparity when there is a dependency between clients and sensitive attributes (scenario 2; $\alpha = 0.9$). Masked Disparity is dominant with high sensitive attribute synergy level across clients. *(middle)* Illustrates PID for varying levels of sensitive attribute heterogeneity (α; see details in Appendix E.6). When α is close to 0.3, the data is split evenly across clients (note $\Pr(Z{=}0){=}0.33$ for the Adult dataset), resulting in a higher level of Unique Disparity. As α deviates from 0.3, i.e., higher dependency between Z and S, the Unique Disparity decreases while Redundant and Masked Disparity increases. *(right)* Illustrates relationship between the synergy level (λ; see details in Appendix E.6) and global and local fairness. As the synergy level increases, the Masked and Local Disparity increases as expected.

- *PID Under Varying Sensitive Attribute Heterogeneity Level.* We partition the dataset across two clients with varying levels of sensitive attribute heterogeneity. We use $\alpha = \Pr(Z{=}0|S{=}0){=}1/4$ to control the sensitive attribute heterogeneity level across clients. Our results are summarized in Fig. 6.5 and Table E.2 in Appendix E.6.

- *Observing Levels of Masked Disparity.* We partition the dataset with varying sensitive attribute synergy levels across clients to study the impact on the Masked Disparity. The *synergy level* $\lambda \in [0, 1]$ measures how closely the true label Y aligns with $Z \oplus S$ (see Definition E.6 in Appendix E.6). Results are summarized in Fig. 6.5 and Table E.3 in Appendix E.6.

- *Experiments Involving Multiple Clients.* We experiment with multiple clients $K = 5$ and $K = 10$. Our findings are presented in Fig. E.2, Fig. E.1 and Table E.4 in Appendix E.6.

6.7 Discussion

Our information-theoretic framework provides a nuanced understanding of the sources of disparity in FL, namely, Unique, Redundant, and Masked disparity. Our experiments offer insights into the agreement and disagreement between local and global fairness under various data distributions. Our experiments and theoretical results show that depending on the data distribution achieving one can often come at the cost of the other (disagreement). The nature of the data distribution across clients significantly impacts the disparity that dominates. Our optimization framework establishes the accuracy-fairness tradeoffs for a dataset and client distribution.

Importantly, our research can: *(i)* inform the use of local disparity mitigation techniques and their convergence and effectiveness when deployed in practice; and *(ii)* also serve as a valuable tool for policy decision-making, shedding light on the effects of model bias at both the global and local levels. This is particularly relevant in the expanding field of algorithmic fairness auditing [24, 288]. Future studies could also investigate how this approach could be extended to more sophisticated fairness measures. The estimation of PID terms largely depends on *(i)* the empirical estimators of the probability distributions; and *(ii)* the efficiency of the convex optimization algorithm used for calculating the unique information. As the number of clients or sensitive attributes increases, the computational cost may rise accordingly. Future work could explore alternative efficient PID computation techniques [280, 281, 289, 290].

Chapter 7: Explaining Group Fairness Trade-offs and Impossibilities

7.1 Introduction

The increasing adoption of machine learning (ML) in high-stakes applications such as employment, finance, healthcare, etc, promises enhanced efficiency and improved decision-making. However, this widespread reliance on ML systems has escalated concerns about the disparate impact [4–7, 24, 248, 267, 291, 292] that these systems might cause on certain *groups*. Several anti-discrimination laws and ethical principles [291] are being actively put forth to ensure algorithmic fairness.

Existing literature has studied a plethora of definitions, metrics, and scholarly debates about algorithmic fairness [4, 5]. Central to the debate of quantifying fairness at a group level are three popular definitions, namely, statistical parity, equalized odds, and predictive parity [4, 5, 246]. Due to the multitude of fairness definitions available, it is often unclear which measure of fairness is most appropriate to adopt in a given setting [293]. Furthermore, it is also well-known that simultaneous satisfiability of these three fairness definitions is generally impossible [294–296].

Given such a fundamental impossibility, practitioners often strive for approximate fairness solutions rather than stringent satisfiability of all these definitions. Such approximate fairness solutions consist of two pivotal aspects: (i) quantification of (un)fairness (i.e., a gap from exact satisfiability); and (ii) development of strategies to mitigate such unfairness in ML models. For instance, one may jointly minimize one or more measures of unfairness while training an ML model which has often led

to empirical tradeoffs between accuracy and different measures of unfairness [256, 257].

Although previous studies have identified certain impossibilities among fairness notions [294–296], a detailed analysis focusing on *the interrelationships among different measures of unfairness*, specifically explaining when they will be in agreement and when they will be in disagreement leading to potential tradeoffs has received limited attention.

Our research bridges this gap by leveraging Partial Information Decomposition (PID) [284], a body of work in information theory, to elucidate the exact relationship between different measures of unfairness. In particular, we consider information-theoretic quantifications [265] of the respective gaps from statistical parity, equalized odds, and predictive parity as our measures of unfairness. Using PID, we demonstrate the exact relationship between these three measures of unfairness in Proposition 7.1. We also refer to Fig. 7.1 and Fig. 7.3 for an illustration of the relationship between the measures.

PID enables us to provide a unified information-theoretic framework that is instrumental in establishing the fundamental limits and tradeoffs among these unfairness measures, particularly in the context of approximate fairness solutions when exact satisfiability of all three fairness definitions is not met. Furthermore, the impossibility among the three fairness definitions can also be derived from our result (see Theorem 7.1). We also identify and delineate the regions of agreement and disagreement among these three measures of unfairness (see Section 7.4.1), providing insights on when there will be a tradeoff and when there will be no tradeoff among the measures of unfairness. We perform numerical simulations on the Adult dataset [37] to complement our theoretical results. Moreover, our work holds broader implications in fields such as algorithmic fairness auditing [24], where it can significantly contribute to the evaluation of fairness in ML models.

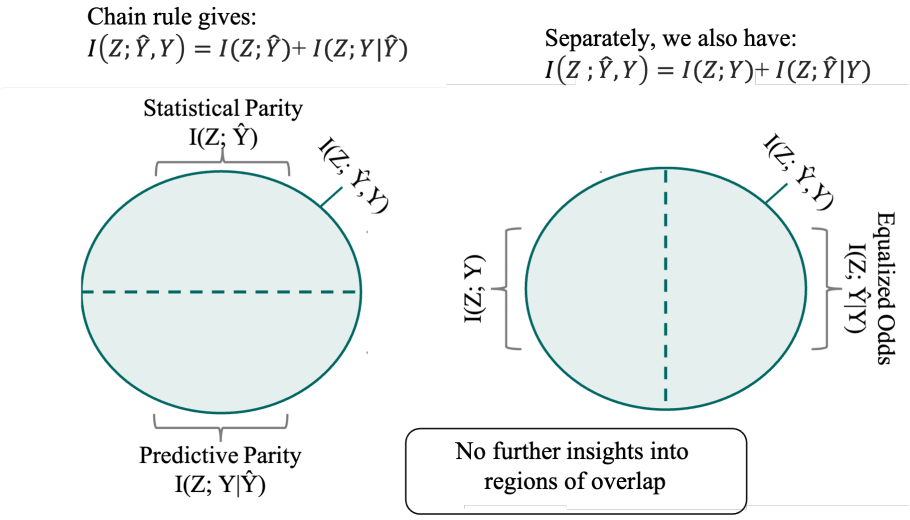


Figure 7.1: Illustrates the decomposition of mutual information $\mathrm{I}(Z;\hat{Y},Y)$ using the chain rule. (*left*) shows the decomposition into Statistical Parity and Predictive Parity. (*right*) shows the decomposition into $\mathrm{I}(Z;Y)$ and Equalized Odds. No further insights into the overlapping regions of these measures, highlighting the need for measures to capture the nuanced interactions between fairness measures.

7.2 Related Works

Existing literature has introduced several foundational definitions and methods for fairness in machine learning (see some comprehensive overviews and surveys [4–7]). Another line of work focuses on exploring trade-offs between fairness and accuracy [21, 255–263, 297–300].

Early works on impossibility results among different fairness measures highlight the challenges of simultaneously satisfying multiple group fairness criteria [294–296, 301, 302]. Notably, [303] challenges the practical implications of the impossibility theorem, showing that fairness across multiple criteria is more achievable than previously believed. Similarly, [300] presents an integer-programming-based framework for optimizing post-processing methods to simultaneously satisfy multiple fairness criteria under small violations while maintaining a minimal reduction in model performance. However, the nuanced interrelationships among different measures of unfairness, specifically explaining

when they will be in agreement and when they will be in disagreement leading to potential tradeoffs has received limited attention.

Information-theoretic measures have been used to quantify group fairness in existing literature [264–270, 272, 273, 304–306]. For instance, [265, 267] have already used mutual information and conditional mutual information to quantify statistical parity, equalized odds, and predictive parity. Closely related to our work, [299] explores whether fairness measures can be gradually compatible using these information-theoretic quantifications and chain rule, also showing that some fairness criteria can be simultaneously improved through fairness-regularized predictors. Our work dives deeper into this nuanced tradeoff among the three group fairness measures leveraging a body of work in information theory called Partial Information Decomposition (PID) that goes beyond chain rule, delineating regions of agreement and disagreement.

PID is recently gaining traction across various ML applications [21, 22, 266, 267, 276, 277, 279, 280, 307, 308]. It is particularly noteworthy in the realm of algorithmic fairness [21, 22, 266, 267]. Prior work [266, 267] leverages PID to dissect total disparity in decision-making into exempt and non-exempt components depending on which features they came from. Another work [21] leverages PID to analyze the trade-offs between global and local fairness in a federated learning setting, identifying three sources of unfairness, and formulates a convex optimization problem to define the theoretical limits of accuracy and fairness trade-offs. We also refer to [22] for a survey of PID in fairness and explainability. Understanding tradeoffs and agreement disagreement between the three canonical group fairness measures, namely statistical parity, equalized odds, and predictive parity, using PID has not been studied. We aim to develop a unified information-theoretic framework that effectively formalizes the fundamental limits and trade-offs among these three unfairness measures.

7.3 Preliminaries

Let X denote the input features, Z denote the sensitive attribute, and Y denote the true label. The sensitive attribute Z is assumed to be binary. We also let $\hat{Y}$ represent the predictions of a model, i.e., $\hat{Y} = f_\theta(X)$ where the model is parameterized by θ. Standard machine learning aims to minimize the empirical risk:

$$\min_\theta L(\theta) = \min_\theta \frac{1}{n} \sum_{i=1}^{n} l(f_\theta(x_i), y_i), \tag{7.1}$$

where $l(\cdot, \cdot)$ is a predefined loss function, x_i is the input feature, $y_i \in \{0, 1\}$ is the true label, and n is the number of datapoints in the dataset. For refer to Section 6.4 for a brief background on Partial information decomposition.

Operational meaning of Unique Information from Blackwell sufficiency: Unique information is closely tethered to Blackwell Sufficiency [309] in statistical decision theory. The concept of Blackwell sufficiency [309] from statistical decision theory helps characterize if a random variable A is more informative than B about Z (also relates to stochastic degradation of channels [308, 310]). A channel $P_{B|Z}$ is Blackwell sufficient with respect to another channel $P_{A|Z}$ (also denoted as $B \geq_Z A$) if there exists a stochastic transformation $P_{A'|B}$ such that the effective channel from Z to A' is equivalent to the original channel from Z to A (see Fig. 7.2). The unique information $\mathrm{Uni}(Z{:}A|B)$ is 0 if and only if $P_{B|Z}$ is Blackwell sufficient with respect to $P_{A|Z}$ [284, 308, 310]. Otherwise, $\mathrm{Uni}(Z{:}A|B) > 0$, and it is viewed as a departure from Blackwell sufficiency, i.e., *there exists a scenario where A gives something unique about Z that you can never get after degrading to B.*

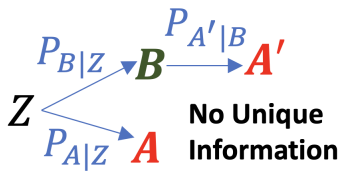


Figure 7.2: Blackwell sufficiency of channel $P_{B|Z}$ with respect to $P_{A|Z}$ means A has no unique information about Z that is not in B.

7.4 Main Contributions

7.4.1 Decomposition of the Measures of Unfairness

We first introduce the information-theoretic quantification corresponding to the three definitions of fairness, namely, statistical parity, equalized odds, and predictive parity. Statistical parity (*independence*), requires the model prediction $\hat{Y}$ to be statistically independent of the sensitive attribute Z. Several measures have been proposed to quantify the gap from statistical parity [5, 242] (essentially dependence between $\hat{Y}$ and Z). In this work, we use the information-theoretic quantification of the statistical parity gap as defined next.

Definition 7.1 (Statistical Parity Gap)**.** *The statistical parity gap of a model f_θ with respect to Z is defined as $I(Z;\hat{Y})$, the mutual information between Z and $\hat{Y}$ (where $\hat{Y} = f_\theta(X)$).*

The concept of statistical parity has often been criticized for not considering the true labels. A perfect predictor $\hat{Y} = Y$ might not satisfy this criterion if Y is correlated to the sensitive attribute Z. Hence, the concept of equalized odds emerges as an alternative definition of fairness [246]. Equalized odds (*separation*) require the model's predictions $\hat{Y}$ to be independent of the sensitive attribute Z, conditioned on the true label Y, i.e., $Z \perp\!\!\!\perp \hat{Y}|Y$.

Definition 7.2 (Equalized Odds Gap)**.** *The equalized odds gap of a model f_θ with respect to Z is defined as $\mathrm{I}(Z;\hat{Y}|Y)$, the conditional mutual information between Z and $\hat{Y}$ given Y.*

Yet another vital fairness measure is predictive parity (*sufficiency*), which focuses on error parity among individuals given the same prediction [4]. Predictive parity requires the sensitive attribute Z to be independent of the true label Y conditioned on the model prediction $\hat{Y}$, i.e., $Z \perp\!\!\!\perp Y|\hat{Y}$.

Definition 7.3 (Predictive Parity Gap)**.** *The predictive parity gap of a model f_θ with respect to Z is defined as $\mathrm{I}(Z;Y|\hat{Y})$, the conditional mutual information between Z and Y given $\hat{Y}$.*

We leverage PID to derive exact relationships among the three measures of unfairness. We decompose the statistical parity gap $I(Z;\hat{Y})$, equalized odds gap $I(Z;\hat{Y}|Y)$, and predictive parity gap $I(Z;Y|\hat{Y})$ into *nonnegative* overlapping terms. The significance of this decomposition is that it highlights regions where these measures are in agreement and disagreement. Fig. 7.1 and Fig. 7.3 provides a pictorial illustration of the overlaps between these three measures of unfairness.

Proposition 7.1. *The statistical parity gap $\mathrm{I}(Z;\hat{Y})$, equalized odds gap $\mathrm{I}(Z;\hat{Y}|Y)$, and predictive parity gap $\mathrm{I}(Z;Y|\hat{Y})$ can be decomposed into nonnegative terms as follows:*

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y},Y). \tag{7.2}$$

$$\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Syn}(Z{:}\hat{Y},Y). \tag{7.3}$$

$$\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y). \tag{7.4}$$

The term $\mathrm{Uni}(Z{:}\hat{Y}|Y)$ quantifies the unique information about the sensitive attribute Z in the model prediction $\hat{Y}$ that is not there in the true label Y. $\mathrm{Uni}(Z{:}\hat{Y}|Y)$ is the common region between the statistical parity gap and the equalized odds gap, highlighting the region where they overlap. The term $\mathrm{Red}(Z{:}\hat{Y},Y)$ quantifies the information about sensitive attribute Z that is common between prediction $\hat{Y}$ and true label Y. $\mathrm{Red}(Z{:}\hat{Y},Y)$ contributes only to the statistical parity gap $\mathrm{I}(Z;\hat{Y})$ and

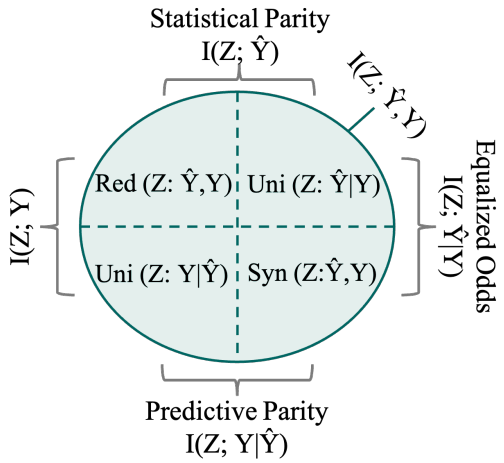


Figure 7.3: Venn diagram showing the exact relationship between the various unfairness measures using PID: A critical observation is that all four PID terms are nonnegative. This enables us to derive several fundamental limits and tradeoffs among the unfairness measures, providing a nuanced understanding of when they agree and disagree.

not to any other measure of unfairness. The term $\mathrm{Syn}(Z{:}\hat{Y}, Y)$ represents the synergistic information about sensitive attribute Z that is *not* present in either $\hat{Y}$ or Y individually but is present jointly in $(\hat{Y}, S)$. $\mathrm{Syn}(Z{:}\hat{Y}, Y)$ is the common region between equalized odds gap and predictive parity gap, highlighting their region of agreement. The unique information $\mathrm{Uni}(Z{:}Y|\hat{Y})$ contributes exclusively to the predictive parity gap $I(Z; Y|\hat{Y})$. This decomposition delineates the distinct regions where these unfairness measures overlap and diverge, offering a nuanced perspective on the interplay in machine learning models.

To better illustrate this decomposition, we now provide examples to understand each of these regions separately. Consider a scenario featuring binary sensitive attributes and true labels i.e., $\hat{Y}, Z, Y \in \{0, 1\}$ with $Z \sim$ Bern(1/2).

Example 7.1 (Pure Uniqueness to Model Prediction)**.** *Let $\hat{Y} = Z$ and $Z \perp\!\!\!\perp Y$ (equal base rate for both groups). Suppose, the model only approves ($Z = 1$) but rejects ($Z = 0$). This model violates both statistical parity and equalized odds, i.e., $\mathrm{I}(Z; \hat{Y}) = \mathrm{I}(Z; \hat{Y}|Y) = 1$. This model satisfies predictive*

parity criterion, $\mathrm{I}(Z;Y|\hat{Y}) = 0$*. This is a case of purely unique information in the model prediction that is not in the true label since all the information about* Z *is derived from the model predictions; the true label* Y *does not correlate with* Z*. Here,* $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 1$*,* $\mathrm{Red}(Z{:}\hat{Y}, Y) = 0$*,* $\mathrm{Syn}(Z{:}\hat{Y}, Y) = 0$*, and* $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0$*.*

Example 7.2 (Pure Redundancy)**.** *Let* $\hat{Y} = Y$ *and* $Y = Z$ *with probability* 0.9*. There is a correlation between the true label* Y *and* Z*, but this model has perfect accuracy. Such a model satisfies equalized odds and predictive parity criterion, i.e.,* $\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{I}(Z;Y|\hat{Y}) = 0$*. However, the model fails to satisfy statistical parity since* $\mathrm{I}(Z;\hat{Y}) = 0.53$*. This is a case of purely redundant information since the information about* Z *is entirely common between both* $\hat{Y}$ *and* Y*. Here,* $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0$*,* $\mathrm{Red}(Z{:}\hat{Y}, Y) = 0.53$*,* $\mathrm{Syn}(Z{:}\hat{Y}, Y) = 0$*, and* $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0$*.*

Example 7.3 (Pure Synergy)**.** *Let* $\hat{Y} = Z$ *XNOR* Y *and* $Z \perp\!\!\!\perp Y$*. The model approves candidates from the group (*$Z = 1$*) with true label* $Y = 1$*, and also from group (*$Z = 0$*) with* $Y = 0$*. On the other hand, it rejects candidates from group (*$Z = 0$*) with true label* $Y = 1$*, and group (*$Z = 1$*) with true label* $Y = 0$*. Such a model violates equalized odds (and predictive parity) as it singularly prefers one group within each true label class. Thus,* $\mathrm{I}(Z;\hat{Y}|Y) = 1$*, and* $\mathrm{I}(Z;Y|\hat{Y}) = 1$*. However, it achieves statistical parity since it maintains an equal approval rate for both groups with* $\mathrm{I}(Z;\hat{Y}) = 0$*. This is a case of synergistic information about* Z *that is not observable in either* $\hat{Y}$ *or* Y *individually but is present jointly in* $Y, \hat{Y}$*. Here,* $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0$*,* $\mathrm{Red}(Z{:}\hat{Y}, Y) = 0$*,* $\mathrm{Syn}(Z{:}\hat{Y}, Y) = 1$*, and* $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0$*.*

Example 7.4 (Pure Uniqueness to True Label)**.** *Let* $Y = Z$ *with probability* 0.9 *and* $Z \perp\!\!\!\perp \hat{Y}$*. The true label* Y *is highly correlated to* Z*, but the model prediction* $\hat{Y}$ *is independent of* Z*. This model violates predictive parity (*$\mathrm{I}(Z;Y|\hat{Y}) = 0.53$*) but satisfies statistical parity and equalized odds (*$\mathrm{I}(Z;\hat{Y}) =$

$\mathrm{I}(Z;\hat{Y}|Y) = 0$*). This is a case of unique information about sensitive attributes in the true label that is not in the model prediction. Here,* $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0$*,* $\mathrm{Red}(Z{:}\hat{Y}, Y) = 0$*,* $\mathrm{Syn}(Z{:}\hat{Y}, Y) = 0$*, and* $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0.53$*.*

These examples demonstrate scenarios of pure uniqueness, redundancy, and synergy to help us understand the decomposition. PID serves as a tool to highlight regions of agreement and disagreement between these fairness definitions. In contrast, traditional fairness metrics lack the granularity to capture these nuanced interactions, making PID an essential asset for a more comprehensive understanding and mitigation of disparities.

We can go beyond the impossibility between the three fairness definitions and further analyze their interrelationships.

Theorem 7.1 (Revisiting Impossibility)**.** *If* $\mathrm{I}(Z;\hat{Y}, Y){>}0$*, at least one of the PID terms, namely,* $\mathrm{Uni}(Z{:}\hat{Y}|Y)$*,* $\mathrm{Red}(Z{:}\hat{Y}, Y)$*,* $\mathrm{Syn}(Z{:}\hat{Y}, Y)$*, or* $\mathrm{Uni}(Z{:}Y|\hat{Y})$ *will be nonnegative. Hence, at least one of the fairness measures, namely, the Statistical Parity Gap (*$\mathrm{I}(Z;\hat{Y})$*), Equalized Odds Gap (*$\mathrm{I}(Z;\hat{Y}|Y)$*), or Predictive Parity Gap (*$\mathrm{I}(Z;Y|\hat{Y})$*) will be nonzero. Conversely, all these unfairness measures will be zero if and only if* $\mathrm{I}(Z;\hat{Y}, Y) = 0$*.*

Proof Sketch: The proof relies on the nonnegativity of each of the PID terms (also recall Fig. 7.3). PID of $\mathrm{I}(Z;\hat{Y}, Y)$ is expressed as $\mathrm{I}(Z;\hat{Y}, Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y}, Y) + \mathrm{Syn}(Z{:}\hat{Y}, Y)$. Since each component in this decomposition is nonnegative, the presence of mutual information ($\mathrm{I}(Z;\hat{Y}, Y) > 0$) implies that at least one of these terms is nonzero. According to Proposition 7.1, each of these PID terms influences at least one unfairness measure. Therefore, the nonnegativity of any one of these terms results in at least one of the unfairness measures being nonzero. □

This is a general result from which one can also derive the impossibility of the three fairness

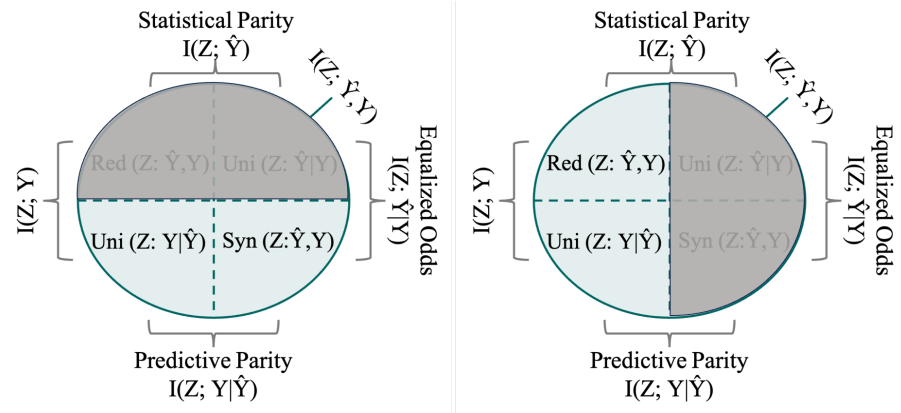


Figure 7.4: (*left*) Illustrates Theorem 7.3, showing that when Statistical Parity is satisfied, the Predictive Parity gap is greater than or equal to the Equalized Odds gap, and if $\mathrm{I}(Z;Y)=0$, then $\mathrm{I}(Z;Y|\hat{Y})=\mathrm{I}(Z;\hat{Y}|Y)$. (*right*) visualizes Theorem 7.5 illustrating that when Equalized Odds is satisfied and $\mathrm{I}(Z;Y)>0$, there is an inverse relationship (trade-off) between Statistical Parity and Predictive Parity ($\mathrm{I}(Z;\hat{Y})=\mathrm{I}(Z;Y)-\mathrm{I}(Z;Y|\hat{Y})$) since $\mathrm{I}(Z;Y)$ is fixed.

definitions under specific conditions. Our next result examines the unfairness measures only when $\mathrm{I}(Z;Y)>0$. It is important to note that $\mathrm{I}(Z;Y)$ is an inherent characteristic of the dataset alone and hence it is independent of the model predictions.

Theorem 7.2 (Dataset Dependent Relationships)**.** *If* $\mathrm{I}(Z;Y)>0$*, either the Statistical Parity Gap* $\mathrm{I}(Z;\hat{Y})$ *or the Predictive Parity Gap* $\mathrm{I}(Z;Y|\hat{Y})$ *must be greater than zero.*

Proof Sketch: The proof relies on demonstrating that the mutual information between Z and Y can be expressed as:

$$\mathrm{I}(Z;Y)=\mathrm{Uni}(Z{:}Y|\hat{Y})+\mathrm{Red}(Z{:}Y,\hat{Y}). \tag{7.5}$$

Though, the PID terms $\mathrm{Uni}(Z{:}Y|\hat{Y})$ and $\mathrm{Red}(Z{:}Y,\hat{Y})$ may vary based on the model chosen, their sum remains constant, reflecting the fixed nature of the mutual information between Z and Y in the dataset. Notably, $\mathrm{Uni}(Z{:}Y|\hat{Y})$ contributes to the predictive parity gap, and $\mathrm{Red}(Z{:}Y,\hat{Y})$ contributes to the statistical parity gap (recall Fig. 7.3). □

7.4.2 Tradeoffs Between Unfairness Measures

In this section, we delineate the fundamental limits and tradeoffs between various unfairness measures. Our findings underscore the intricate and sometimes conflicting nature of different fairness objectives in algorithmic decision-making. Examining fairness through the lens of PID uncovers the nuanced interplay between different unfairness measures.

First, we will explore scenarios where models are trained with a focus on achieving any one specific fairness criterion and analyze its implications on the other two fairness notions. This applies to models that have been trained to achieve fairness either through in-processing techniques, such as adding fairness regularizers to the loss function, or through post-processing methods that adjust model outputs after training.

Theorem 7.3. *If Statistical Parity is satisfied, i.e.,* $\mathrm{I}(Z;\hat{Y}) = 0$*, then the Predictive Parity Gap is greater than the Equalized Odds Gap, i.e.,* $\mathrm{I}(Z;Y|\hat{Y}) \geq \mathrm{I}(Z;\hat{Y}|Y)$*. Additionally, if the dataset is such that* $\mathrm{I}(Z;Y) = 0$*, then Predictive Parity and Equalized Odds are equivalent, i.e.,* $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{I}(Z;\hat{Y}|Y)$.

Proof Sketch: We refer to Fig. 7.3 for an intuitive understanding of the proof. Given that Statistical Parity is zero, we have $\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y},Y) = 0$. Since all PID terms are non-negative, it follows that individually $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0$ and $\mathrm{Red}(Z{:}\hat{Y},Y) = 0$. Consequently, the Equalized Odds gap simplifies to $\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{Syn}(Z{:}\hat{Y},Y)$. On the other hand, the Predictive Parity gap is $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y)$. Since all PID terms are nonnegative, it follows that $\mathrm{I}(Z;Y|\hat{Y}) \geq \mathrm{I}(Z;\hat{Y}|Y)$.

Furthermore, when $\mathrm{I}(Z;Y) = 0$, it results in $\mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y},Y) = 0$, leading to each of those individual terms being zero, i.e., $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0$ and $\mathrm{Red}(Z{:}\hat{Y},Y) = 0$. Therefore,

$\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{I}(Z;\hat{Y}|Y)$ (see Fig. 7.4 for an illustration). □

Similar to Theorem 7.3, one can also derive the relationship between the statistical parity gap and equalized odds gap when predictive parity is satisfied.

Theorem 7.4. *If Predictive Parity is satisfied, i.e.,* $\mathrm{I}(Z;Y|\hat{Y})=0$*, then the Statistical Parity Gap is greater than the Equalized Odds Gap, i.e.,* $\mathrm{I}(Z;\hat{Y}) \geq \mathrm{I}(Z;\hat{Y}|Y)$*. Additionally, if the dataset is such that* $\mathrm{I}(Z;Y) = 0$*, then Statistical Parity and Equalized Odds are equal, i.e.,* $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{I}(Z;\hat{Y}|Y)$.

Theorem 7.3 and 7.4 demonstrate scenarios where unfairness measures are in agreement, now we provide a third scenario where two measures of unfairness will be in disagreement.

Theorem 7.5. *If Equalized Odds is satisfied, i.e.,* $\mathrm{I}(Z;\hat{Y}|Y)=0$ *and* $\mathrm{I}(Z;Y) > 0$*, an inverse relationship (tradeoff) exists between Statistical Parity and Predictive Parity, i.e.,* $\mathrm{I}(Z;\hat{Y}) = \mathrm{I}(Z;Y) - \mathrm{I}(Z;Y|\hat{Y})$*. Thus, increasing one leads to a decrease in the other, and vice versa.*

Proof Sketch Given that Equalized Odds is met, we have $\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Syn}(Z{:}\hat{Y},Y) = 0$. Consequently, from nonnegativity, both the terms $\mathrm{Uni}(Z{:}\hat{Y}|Y)$ and $\mathrm{Syn}(Z{:}\hat{Y},Y)$ are 0. Statistical Parity gap simplifies to $\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y},Y) = \mathrm{Red}(Z{:}\hat{Y},Y)$, and the Predictive Parity gap is expressed as $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{Uni}(Z{:}Y|\hat{Y})$. Hence, $\mathrm{I}(Z;Y) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y},Y) = \mathrm{I}(Z;\hat{Y}) + \mathrm{I}(Z;Y|\hat{Y})$. Since, $\mathrm{I}(Z;Y)$ is fixed for a dataset, an increase in the statistical parity gap leads to a decrease in the predictive parity gap, and vice versa (see Fig. 7.4 for an illustration). □

7.5 Experiments

In this section, we provide an experimental demonstration on the Adult dataset [37] to validate our theoretical findings. The classification task for this dataset involves predicting whether an individ-

Table 7.1: Results of Regularizers on Different Measures of Unfairness

Regularizers	Statistical Parity $\mathrm{I}(Z;\hat{Y})$	Equalized Odds $\mathrm{I}(Z;\hat{Y}\|Y)$	Predictive Parity $\mathrm{I}(Z;Y\|\hat{Y})$	
	$\mathrm{Red}(Z{:}\hat{Y},Y)$	$\mathrm{Uni}(Z{:}\hat{Y}\|Y)$	$\mathrm{Syn}(Z{:}\hat{Y},Y)$	$\mathrm{Uni}(Z{:}Y\|\hat{Y})$
SP	0.012	0.000	0.001	0.024
PP	0.026	0.007	0.008	0.011
EO	0.011	0.000	0.001	0.026
EO, PP	0.000	0.000	0.000	0.037
SP, PP	0.000	0.000	0.000	0.037
SP, EO	0.008	0.000	0.000	0.028
SP, EO, PP	0.000	0.000	0.000	0.037

ual's income exceeds 50K, using features such as occupation, education, etc.

We train a neural network consisting of a sequence of layers: the input layer is followed by three hidden layers, each with 32 units and ReLU activation, and concludes with a single output layer using a sigmoid activation function. Training is conducted using a batch size of 512, and the Adam optimizer with a learning rate of 0.01. We apply various fairness regularizers and measure the unfairness as well as their decomposition (results are summarized in Table.7.1). We use the *dit* package [287] for PID computation and *FairTorch* [311] for fairness regularizer implementation.

A key observation in our analysis is that $\mathrm{I}(Z;Y)$ consistently measures 0.037 using the Adult dataset. This mass does not decrease across various models since it only depends on the dataset. The PID terms in $\mathrm{I}(Z;Y)$, i.e., $\mathrm{Uni}(Z{:}Y|\hat{Y})$ and $\mathrm{Red}(Z{:}\hat{Y},Y)$ contribute to either predictive parity or statistical parity gap. When statistical parity is achieved (scenario with SP regularizer), the predictive parity gap is greater than the equalized odds gap. Also due to the impossibility of attaining zero unfairness with all the measures (see scenario with SP, EO, and PP regularizers), the mass typically moves to $\mathrm{Uni}(Z{:}Y|\hat{Y})$, contributing to the predictive parity. The experiments confirm the theoretical insights by illustrating the inherent trade-offs between fairness measures. Specifically, even with fairness regu-

larizers, achieving zero unfairness across all metrics is impossible due to the fixed information content in the dataset, i.e., $\mathrm{I}(Y;Z)$. The movement of unfairness mass between statistical parity and predictive parity highlights the difficulty of balancing fairness across multiple criteria, reinforcing the necessity of carefully considering these trade-offs in real-world applications.

7.6 Discussion

By introducing this unifying framework, we provide a tool for gaining a more nuanced understanding of the interplay between different unfairness measures, thereby improving the decision-making and deployment of fair machine learning systems. Our work holds broader implications in fields such as fairness auditing [24], explainability [16, 312, 313], policy regulation [291, 296], where it can significantly contribute to the evaluation and understanding of unfairness in machine learning models. This work not only furthers the theoretical discourse but would also have significant societal implications, guiding the trajectory toward more responsible and equitable machine learning in high-stakes settings. Future work could explore efficient methods to estimate PID [280, 290, 307, 314, 315] (also see [22] for more discussion on PID estimation).

Appendix A

A.1 Relevant Inequalities

Lemma A.1 (Cauchy-Schwarz Inequality)**.** *If* $\mathbf{u}, \mathbf{v} \in V$*, where* V *is a vector space, then*

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|.$$

This inequality is an equality if and only if one of $\mathbf{u}, \mathbf{v}$ *is a scalar multiple of the other.*

Lemma A.2 (Sedrakyan's Inequality)**.** *Let* $u_1, u_2, \ldots, u_n$ *and* $v_1, v_2, \ldots, v_n$ *be positive real numbers. Then:*

$$\left(\frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} v_i} \right)^2 \leq \sum_{i=1}^{n} \frac{u_i^2}{v_i^2}$$

This inequality is a direct consequence of the Cauchy–Schwarz inequality (see Lemma A.1), obtained by taking the dot product in $\mathbb{R}^n$ *upon substituting* $u_i' = u_i / \sqrt{v_i}$ *and* $v_i' = \sqrt{v_i}$*.*

Lemma A.3 (Jensens Inequality)**.** *Let* X *be an integrable random variable. Let* $g : \mathbb{R}^d \to \mathbb{R}$ *be a convex function such that* $Y = g(X)$ *is also integrable. Then, the following inequality, called Jensen's inequality, holds:*

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X]).$$

A.2 Proof of Theorem 2.1

Theorem 2.1 (Connection to Rashomon Effect)**.** *For points $x_1, \ldots, x_n \in \mathcal{X}$ (lying on the data-manifold) under naturally-occurring model change, the following holds:*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}|M(x_i) - m(x_i)|\right] \leq \sqrt{\nu}, \quad \textit{where} \quad \nu = \frac{1}{n}\sum_{i=1}^{n}\nu_{x_i} \tag{2.1}$$

Proof.

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}|m(x_i) - M(x_i)|\right] \overset{(a)}{\leq} \mathbb{E}\left[\sqrt{\sum_{i=1}^{n}\frac{1}{n^2}}\sqrt{\sum_{i=1}^{n}|m(x_i) - M(x_i)|^2}\right] \tag{A.1}$$

$$= \frac{1}{\sqrt{n}}\,\mathbb{E}\left[\sqrt{\sum_{i=1}^{n}|m(x_i) - M(x_i)|^2}\right] \tag{A.2}$$

$$\overset{(b)}{\leq} \frac{1}{\sqrt{n}}\sqrt{\mathbb{E}\left[\sum_{i=1}^{n}|m(x_i) - M(x_i)|^2\right]} \tag{A.3}$$

$$\overset{(c)}{=} \frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}\nu_{x_i}} = \sqrt{\nu} \tag{A.4}$$

$$\tag{A.5}$$

Here (a) holds from Cauchy-Schwarz Inequality (Lemma A.1) applied on the dot product of the two vectors $[1/n, 1/n, \ldots, 1/n]$ and $[|m(x_1)-M(x_1)|, |m(x_2)-M(x_2)|, \ldots, |m(x_n)-M(x_n)|]$. Next, (b) holds from Jensen's Inequality (Lemma A.3) applied on concave function $f(u) = \sqrt{u}$. Finally, (c) holds because the points $x_1, x_2, \ldots, x_n$ lie on the data-manifold and hence the variance $\nu_{x_i} \leq \nu$ from Definition 2.5.

□

A.3 Proof of Theorem 2.2

Theorem 2.2 (Impossibility Under Adversarial Change)**.** *Given a model and counterfactual, one can always design a similar model such that the particular targeted counterfactual can be invalidated.*

Proof. We aim to construct a new model M such that M is "similar" to m on the entire input space $\mathcal{X}$ except at the point x', where it inverts the prediction at x'. Define the new model M as follows:

$$M(x) = \begin{cases} 1 - m(x') & \text{if } x = x', \\ m(x) & \text{otherwise.} \end{cases}$$

This construction ensures that M behaves exactly like m for all $x \in \mathcal{X} \setminus \{x'\}$, maintaining the "similarity" between M and m. However, at the point x', the outcome is explicitly set to be the inverse of $m(x')$, i.e., $M(x') = 1 - m(x')$, thereby invalidating the counterfactual instance x' under new model M. This demonstrates that for a given model, one can design another similar model such that any particular targeted counterfactual can be invalidated. □

A.4 Proof of Theorem 2.3

To prove Theorem 2.3, we begin with Lemma 2.1. Assume the changed model M comes from a discrete class of random variables. A possible realization of a model is denoted by $\tilde{m}_i$, with $i = 1, 2, \ldots, n$. The set of all possible models is denoted by $\mathcal{M} = \{\tilde{m}_1, \tilde{m}_2, \ldots, \tilde{m}_n\}$. Let $M = \tilde{m}_i$ with probability p_i such that $\sum_{i=1}^{n} p_i = 1$.

Lemma 2.1 (Deviation Bound)**.** *Let* $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$, $Z = \frac{1}{k} \sum_{i=1}^{k} (m(X_i) - M(X_i))$, *and*

$|\mathbb{E}[Z|M] - \mathbb{E}[Z]| < \epsilon'$. *Then, under naturally-occurring change,* $\mathbb{E}[Z] = 0$. *Moreover, for any* $\epsilon > 2\epsilon'$,

$$\Pr(Z \geq \epsilon) \leq \exp\left(\frac{-k\epsilon^2}{8(\gamma + \gamma_m)^2\sigma^2}\right). \tag{2.3}$$

Proof. To prove Lemma 2.1, notice that,

$$\begin{aligned}\mathbb{E}[Z] &= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}_{X_i}\left[\mathbb{E}_{M|X_i}[m(X_i) - M(X_i)]\right] \\ &\overset{(a)}{=} \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}_{X_i}[m(X_i) - m(X_i)] = 0,\end{aligned} \tag{A.6}$$

where (a) holds from the naturally occurring model change assumption (see Definition 1). The remaining part of the proof leverages concentration bounds for Lipschitz functions of independent Gaussian random variables outlined in Lemma 2.2.

Lemma 2.2 (Gaussian Concentration Inequality)**.** *Let* $W = (W_1, W_2, \ldots, W_n)$ *consist of* n *i.i.d. random variables belonging to* $\mathcal{N}(0, \sigma^2)$*, and* $Z = f(W)$ *be a* γ*-Lipschitz function, i.e.,* $|f(W) - f(W')| \leq \gamma\|W - W'\|$. *Then:*

$$\Pr(Z - \mathbb{E}[Z] \geq \epsilon) \leq \exp\left(\frac{-\epsilon^2}{2\gamma^2\sigma^2}\right) \textit{ for all } \epsilon > 0. \tag{2.4}$$

Let X_{ij} denote the j-th element of $X_i \in \mathbb{R}^d$ and x_j denote the j-th element of $x \in \mathbb{R}^d$. We define a $k \times d$ matrix $W = [W_{ij}]_{i=1,2,\ldots,k, \text{ and } j=1,2,\ldots,d}$ with $W_{ij} = X_{ij} - x_j$. Notice that, $W_{ij} \sim \mathcal{N}(0, \sigma^2)$ for

$i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, d$ and, we can write $Z = f(W)$ with Lipschitz constant $(\gamma_m + \gamma)/\sqrt{k}$.

$$
\begin{aligned}
&|f(W) - f(W')| \\
&= \left| \frac{1}{k} \sum_{i=1}^{k} (m(X_i) - M(X_i) - m(X_i') + M(X_i')) \right| \\
&\overset{(a)}{\leq} \frac{1}{k} \sum_{i=1}^{k} (|m(X_i) - m(X_i')| + |M(X_i) - M(X_i')|) \\
&\overset{(b)}{\leq} \frac{1}{k} \sum_{i=1}^{k} (\gamma_m + \gamma) \|X_i - X_i'\|_2 \\
&\overset{(c)}{\leq} \frac{(\gamma_m + \gamma)\sqrt{k}}{k} \sqrt{\sum_{i=1}^{k} \|X_i - X_i'\|_2^2} \\
&= \frac{(\gamma_m + \gamma)}{\sqrt{k}} \sqrt{\sum_{i=1}^{k} \sum_{j=1}^{d} |W_{ij} - W_{ij}'|^2} \\
&= \frac{(\gamma_m + \gamma)}{\sqrt{k}} \|W - W'\|_2. \qquad \text{(A.7)}
\end{aligned}
$$

□

Here, (a) holds from the triangle inequality, (b) follows directly from the definition of γ-Lipschitz (Definition 2.1), and (c) from using the Cauchy-Schwarz Inequality (Lemma A.1).

Now, we substitute these expressions in the Gaussian concentration bound (Lemma 2.2).

$$
\Pr(Z - \mathbb{E}[Z|M = \tilde{m}] \geq \tilde{\epsilon} | M = \tilde{m}) \leq \exp\left(\frac{-k\tilde{\epsilon}^2}{2(\gamma + \gamma_m)^2 \sigma^2} \right). \qquad \text{(A.8)}
$$

Since $|\mathbb{E}[Z|M] - \mathbb{E}[Z]| < \epsilon'$ and $\mathbb{E}[Z] = 0$, we have $-\epsilon' < \mathbb{E}[Z|M = \tilde{m}] < \epsilon' \; \forall \tilde{m} \in \mathcal{M}$. Now observe that,

$$\Pr(Z \geq \epsilon' + \tilde{\epsilon} | M = \tilde{m}) \overset{(a)}{\leq} \Pr(Z \geq \mathbb{E}\left[Z | M = \tilde{m}\right] + \tilde{\epsilon} | M = \tilde{m}) \tag{A.9}$$

$$\leq \exp\left(\frac{-k\tilde{\epsilon}^2}{2(\gamma + \gamma_m)^2\sigma^2}\right). \tag{A.10}$$

Here, (a) holds since $\mathbb{E}\left[Z | M = \tilde{m}\right] < \epsilon'$. The event on the left is a subset of that on the right. Therefore, the probability of the event $\{Z \geq \epsilon' + \tilde{\epsilon}\}$ occurring cannot be more than the probability of the event $\{Z \geq \mathbb{E}\left[Z | M = \tilde{m}\right] + \tilde{\epsilon}\}$ occurring.

$$\Pr(Z \geq \epsilon' + \tilde{\epsilon}) \overset{(b)}{=} \sum_{i=1}^{n} \Pr(Z \geq \epsilon' + \tilde{\epsilon} | M = \tilde{m}_i) \Pr(M = \tilde{m}_i) \tag{A.11}$$

$$\overset{(c)}{\leq} \exp\left(\frac{-k\tilde{\epsilon}^2}{2(\gamma_m + \gamma)^2\sigma^2}\right) \sum_{i=1}^{n} \Pr(M = \tilde{m}_i) \tag{A.12}$$

$$= \exp\left(\frac{-k\tilde{\epsilon}^2}{2(\gamma_m + \gamma)^2\sigma^2}\right) \tag{A.13}$$

$$\overset{(d)}{\leq} \exp\left(\frac{-k(\tilde{\epsilon} + \epsilon')^2}{8(\gamma_m + \gamma)^2\sigma^2}\right) \tag{A.14}$$

Here, (b) holds from the law of total probability. Next, (c) follows from bound in (A.9). Finally, (d) holds from the inequality $4\tilde{\epsilon}^2 > (\tilde{\epsilon} + \epsilon')^2$ which holds for $\tilde{\epsilon} > \epsilon' > 0$. Setting $\epsilon = \tilde{\epsilon} + \epsilon'$ completes the proof of Lemma 2.1.

Theorem 2.3 (Probabilistic Guarantee)**.** *Let $X_1, X_2, \ldots, X_k$ be k iid random variables with distribution $\mathcal{N}(x, \sigma^2 I_d)$ and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Suppose $|\mathbb{E}\left[Z|M\right] - \mathbb{E}\left[Z\right]| < \epsilon'$. Then, for any $\epsilon > 2\epsilon'$, a counterfactual $x \in \mathcal{X}$ under naturally-occurring model change satisfies:*

$$\Pr(M(x) \geq R_{k,\sigma^2}(x, m) - \epsilon) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma_m + \gamma)^2\sigma^2}\right).$$

Probability is over the randomness of both M and X_i's.

Proof. The Lipschitz property of $M(\cdot)$ around x is given by,

$$|M(x) - M(x')| \leq \gamma \|x - x'\| \text{ for all } x, x' \in \mathbb{R}^d$$

Therefore,

$$M(x) \geq M(X_i) - \gamma \|x - X_i\| \tag{A.15}$$

$$M(x) \overset{(a)}{\geq} \frac{1}{k} \sum_{i=1}^{k} (M(X_i) - \gamma \|x - X_i\|) \tag{A.16}$$

where (a) holds from taking the average of the inequality (A.15) over all i from 1 to k.

From Lemma 2.1, for $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$,

$$\frac{1}{k} \sum_{i=1}^{k} M(X_i) \geq \frac{1}{k} \sum_{i=1}^{k} m(X_i) - \epsilon \tag{A.17}$$

with probability at least $1 - \exp\left(-\frac{k\epsilon^2}{8(\gamma+\gamma_m)^2\sigma^2}\right)$.

Hence, plugging (A.17) into (A.16), we have:

$$\Pr\left(M(x) \geq \frac{1}{k} \sum_{i=1}^{k} (m(X_i) - \gamma \|x - X_i\|) - \epsilon\right) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma + \gamma_m)^2\sigma^2}\right).$$

Recall from Definition 2.6, $R_{k,\sigma^2}(x, m) = \frac{1}{k} \sum_{i=1}^{k} (m(X_i) - \gamma \|x - X_i\|)$. Hence, we have:

$$\Pr(M(x) \geq R_{k,\sigma^2}(x, m) - \epsilon) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma + \gamma_m)^2\sigma^2}\right).$$

□

A.5 Proof of Theorem 2.4

Theorem 2.4. *Let $X_1, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$ on the data-manifold and $Z{=}\frac{1}{k}\sum_{i=1}^{k}(m(X_i){-}M(X_i))$. Then, a counterfactual $x \in \mathcal{X}$ under Assumption 2.1 and naturally-occurring model change satisfies:*

$$\Pr(M(x) \geq R_{k,\sigma^2}(x,m) - \epsilon) \geq \left(1 - \frac{\alpha}{\epsilon^2}\right)\left(1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma_m{+}\gamma)^2\sigma^2}\right)\right).$$

Probability is over the randomness of both M and X_i's.

Corollary 2.1. *Let $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$ and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Suppose $\Pr\left(|\mathbb{E}\left[Z|M\right] - \mathbb{E}\left[Z\right]| > \epsilon'\right) \leq \delta$. Then, for any $\epsilon > 2\epsilon'$, a counterfactual $x \in \mathcal{X}$ under naturally-occurring model change satisfies:*

$$\Pr(M(x) \geq R_{k,\sigma^2}(x,m){-}\epsilon) \geq (1-\delta)\left(1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma_m{+}\gamma)^2\sigma^2}\right)\right).$$

Probability is over the randomness of both M and X_i's.

Proof. Let $S = \{\tilde{m}_i : |\mathbb{E}[Z|M = \tilde{m}_i] - \mathbb{E}[Z]| < \tilde{\epsilon}\}$.

$$\begin{aligned}
&\Pr(Z \geq c + (\tilde{\epsilon} + \epsilon)) = \mathbb{E}[\mathbb{1}(Z > c + (\tilde{\epsilon} + \epsilon)] \\
&= \mathbb{E}_M\left[\mathbb{E}_{X_i|M}[\mathbb{1}(Z > c + (\tilde{\epsilon} + \epsilon))]\right] \\
&= \sum_S \mathbb{E}_{X_i|M}[\mathbb{1}(Z > c + (\tilde{\epsilon} + \epsilon))] \Pr(M = \tilde{m}_i) \\
&+ \sum_{M/S} \mathbb{E}_{X_i|M}[\mathbb{1}(Z > c + (\tilde{\epsilon} + \epsilon))] \Pr(M = \tilde{m}_i) \\
&\leq \sum_S \Pr(M = \tilde{m}_i) \exp\left(\frac{-k(\tilde{\epsilon} + \epsilon)^2}{8(\gamma_m + \gamma)^2\sigma^2}\right) + \sum_{M/S} \Pr(M = \tilde{m}_i) \\
&= (1 - \delta) \exp\left(\frac{-k(\tilde{\epsilon} + \epsilon)^2}{8(\gamma_m + \gamma)^2\sigma^2}\right) + \delta.
\end{aligned}$$

Where $\delta = 1 - \Pr(M \in S)$. The remaining part of the proof can be derived from Theorem 2.3. □

Lemma 2.3. *Let $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$ on the data-manifold and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Then, for all $\epsilon > 0$, under naturally-occurring model change and Assumption 2.1,*

$$\Pr(|\mathbb{E}[Z|M] - \mathbb{E}[\mathbb{E}[Z|M]]| \geq \epsilon) \leq \frac{\alpha}{\epsilon^2}. \tag{2.5}$$

Proof. Given a random variable $E[Z|M]$, by Chebyshev's inequality, we have:

$$\Pr\left(|E[Z|M] - \mathbb{E}[E[Z|M]]| \geq \varepsilon\right) \leq \frac{\text{Var}(E[Z|M])}{\varepsilon^2}. \tag{A.18}$$

Let, the set of all possible changed models $\mathcal{M} = \{\tilde{m}_1, \tilde{m}_2, \ldots, \tilde{m}_{|\mathcal{M}|}\}$ and $M = \tilde{m}_i$ with probability

p_j such that $\sum_{j=1}^{|\mathcal{M}|} p_j = 1$. The variance of $E[Z|M]$ is given by:

$$\text{Var}(E[Z|M]) \overset{(a)}{=} \sum_{j=1}^{|\mathcal{M}|} \left(E_{X_i|M} \left[\frac{1}{k} \sum_{i=1}^{k} m(X_i) - \tilde{m}_j(X_i) \right] \right)^2 \Pr(M = \tilde{m}_j) \tag{A.19}$$

$$= \frac{1}{k^2} \sum_{j=1}^{|\mathcal{M}|} \left(\sum_{i=1}^{k} E_{X_i|M} \left[m(X_i) - \tilde{m}_j(X_i) \right] \right)^2 \Pr(M = \tilde{m}_j) \tag{A.20}$$

$$\overset{(b)}{\leq} \frac{1}{k^2} \sum_{j=1}^{|\mathcal{M}|} \left(k \sum_{i=1}^{k} \left(E_{X_i|M} [m(X_i) - \tilde{m}_j(X_i)] \right)^2 \right) \Pr(M = \tilde{m}_j) \tag{A.21}$$

$$= \frac{1}{k} \sum_{j=1}^{|\mathcal{M}|} \sum_{i=1}^{k} \left(\int_{-\infty}^{+\infty} (m(x_i) - \tilde{m}_j(x_i)) f(x_i|\tilde{m}_j) dx_i \right)^2 \Pr(M = \tilde{m}_j) \tag{A.22}$$

$$\overset{(c)}{\leq} \frac{1}{k} \sum_{j=1}^{|\mathcal{M}|} \sum_{i=1}^{k} \int_{-\infty}^{+\infty} (m(x_i) - \tilde{m}_j(x_i))^2 f(x_i|\tilde{m}_j) \, dx_i \Pr(M = \tilde{m}_j) \tag{A.23}$$

$$= \sum_{j=1}^{|\mathcal{M}|} \int_{-\infty}^{+\infty} (m(x) - \tilde{m}_j(x))^2 f(x|\tilde{m}_j) \, dx_i \Pr(M = \tilde{m}_j) \tag{A.24}$$

$$= \sum_{j=1}^{|\mathcal{M}|} E_{X|M}[(m(X) - M(X))^2] \Pr(M = \tilde{m}_j) \tag{A.25}$$

$$= E_M \left[E_{X|M}[(m(X) - M(X))^2] \right] \tag{A.26}$$

$$= E_X \left[E_{M|X}[(m(X) - M(X))^2] \right] \tag{A.27}$$

$$\overset{(d)}{=} E_X[\nu_X] \overset{(e)}{\leq} \alpha \tag{A.28}$$

Here, (a) holds from the definition of variance for conditional expectation, (b) holds from the Sedrakyan Inequality (see Lemma A.2), (c) holds from Jensen's inequality (see Lemma A.3), (d) holds from condition (3) of natural occurring model changes (see Definition 2.5), and (e) holds from Assumption 2.1. □

A.6 Proof of Theorem 2.5

Lemma A.4 (Expected Norm of a Multivariate Gaussian)**.** *Let $X \sim \mathcal{N}(0, \sigma^2 I_d)$, then;*

$$\mathbb{E}\left[\|X\|\right] = \sqrt{2\sigma^2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}.$$

where $\Gamma(z)$ denotes the Gamma function, i.e., $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$.

Theorem 2.5 (Fundamental Lower Bound on Sample Size)**.** *Let $\mathcal{M}$ be a class of all models with Lipschitz constant γ in domain $[-T, T]^d \subset \mathbb{R}^d$ and bound on the second-order partial derivatives, i.e., $\forall m \in \mathcal{M}$, $\left|\frac{\partial^2 m}{\partial x_i \partial x_j}\right| \leq \psi$ for all $x \in \mathbb{R}^d$ and $i, j \in \{1, 2, \ldots, d\}$. If,*

$$\textit{sup}_{m\in\mathcal{M}}\mathbb{E}\left[\left|\hat{R}_{k,\sigma^2}(x, m) - R_{k,\sigma^2}(x, m)\right|\right] < \epsilon,$$

then $k \geq \left(\frac{\sqrt{2\sigma^2}\, T\, \psi\, \Gamma\left(\frac{d+1}{2}\right)}{9.69\epsilon\, \Gamma\left(\frac{d}{2}\right)}\right)^d$, where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ (Gamma function).

Proof. Let $\mathcal{M}$ be a class of all twice differentiable functions with Lipschitz constant γ and bound on the second-order partial derivatives, i.e., $\left|\frac{\partial m}{\partial x_i \partial x_j}\right| \leq \psi$ for all $x \in \mathbb{R}^d$ and $i, j \in \{1, 2, \ldots, d\}$.

To prove Theorem 2.5, we show that there exists a set $\mathcal{M}' \subseteq \mathcal{M}$ such that,

$$\text{sup}_{m\in\mathcal{M}'}\mathbb{E}\left[\left|\hat{R}_{k,\sigma^2}(x, m) - R_{k,\sigma^2}(x, m)\right|\right] > \epsilon \tag{A.29}$$

when $k < \left(\frac{\sqrt{2\sigma^2}\, T\, \psi\, \Gamma\left(\frac{d+1}{2}\right)}{9.69\epsilon\, \Gamma\left(\frac{d}{2}\right)}\right)^d$.

First, we simplify the expression:

$$\sup_{m\in\mathcal{M}'}\mathbb{E}_X\left[\left|\hat{R}_{k,\sigma^2}(x,m)-R_{x,\sigma^2}(x,m)\right|\right] \tag{A.30}$$

$$=\sup_{m\in\mathcal{M}'}\mathbb{E}_X\left[|m(X)-\hat{\gamma}_x\|x-X\|-m(X)+\gamma\|x-X\||\right]$$

$$=\sup_{m\in\mathcal{M}'}\mathbb{E}_X\left[\|x-X\||\hat{\gamma}_x-\gamma|\right] \tag{A.31}$$

This reduces to finding a fundamental lower bound on the sample size needed to precisely estimate the stability measure using queries from the model.

We aim to construct the set $\mathcal{M}'$, such that for all functions g in this set, i.e., $g \in \mathcal{M}'$, the following conditions hold:

$$\max_{x\in\mathbb{R}^d}\|\nabla g(x)\|_2 \geq \epsilon + L,$$
$$\max_{x\in\mathbb{R}^d}\left|\frac{\partial^2 g}{\partial x_k \partial x_j}(x)\right| \leq \psi.$$

To do this, consider a radial (bump) function with radius ρ with center z.

$$g_z(x) = \begin{cases} 1.25\epsilon\rho\exp\left(-\frac{1}{1-\alpha^{-2}\sum_{j=1}^{d}(x_j-z_j)^2}\right)+f_L(x) & \text{if } \|x-z\|_2<\rho \\ f_L(x) & \text{otherwise.} \end{cases} \tag{A.32}$$

where $f_L(x)$ is a linear function with gradient L. This radial function is constructed such that:

$$\max_{x\in\mathbb{R}^d}\|\Delta g_z(x)\|_2 = \epsilon + L. \tag{A.33}$$

$$\max_{x_i,x_j\in\{1,2,\ldots,d\}} \left|\frac{\partial g}{\partial x_i \partial x_j}\right| = 9.96\epsilon/\rho. \tag{A.34}$$

Using the radial function to construct the set $\mathcal{M}'$, $\max_{x\in\mathbb{R}^d,i,j\in\{1,\ldots,d\}} \left|\frac{\partial^2 g}{\partial x_i \partial x_j}(x)\right| \leq \psi$ implies that $\rho = \frac{9.69\epsilon}{\psi}$.

Now consider the number of $\|.\|_2$-hyperspheres $\{B_i\}_{i=1,\ldots,k}$ of radius ρ that fit in the domain $[-T,T]^d \in \mathbb{R}^d$ such that: $\forall B_i, B_j, B_i \cap B_j = \emptyset$ if $i \neq j$. A lower bound on the number of hyperspheres of radius ρ that can fit in a hypercube of length l is $(\frac{l}{2\rho})^d$ hence for the domain $[-T,T]^d$;

$$\left(\frac{2T}{2\rho}\right)^d = \left(\frac{T\psi}{9.69\epsilon}\right)^d. \tag{A.35}$$

The set $\mathcal{M}'$ can be constructed by using the set of ball centers z_i for B_i, i.e.,

$$\mathcal{M}' = \{g_{z_i} \text{ with } \rho = \frac{9.69\epsilon}{\psi}, \forall i = 1, 2, \ldots, k\}.$$

Suppose that $k < \left(\frac{T\psi}{9.69\epsilon}\right)^d$. By construction, there exists a ball B_i with associated ball center (z_i) such that no observations are sampled in B_i. Hence, our Lipschitz constant estimate $\hat{\gamma}_x$ for that function will underestimate the Lipschitz constant ($\hat{\gamma}_x = L$) instead of the true Lipschits constant ($\gamma = \epsilon + L$).

Continuing from Eqn. A.31,

$$\sup_{m\in\mathcal{M}'} \mathbb{E}_X \left[\|x - X\| |\hat{\gamma}_x - \gamma|\right] \geq \sup_{m\in\mathcal{M}'} \mathbb{E}_X \left[\|x - X\| |L - L - \epsilon|\right] \tag{A.36}$$

$$= \epsilon \mathbb{E}_X \left[\|x - X\|\right] \tag{A.37}$$

$$\overset{(a)}{=} \epsilon\sqrt{2\sigma^2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \tag{A.38}$$

Where step (a) follows from the expected norm of a multivariate Gaussian (see Lemma A.4).

Hence, if

$$\sup_{m\in\mathcal{M}}\mathbb{E}_X\left[\left|\hat{R}_{k,\sigma^2}(x,m)-R_{x,\sigma^2}(x,m)\right|\right]<\epsilon', \tag{A.39}$$

then,

$$k\geq\left(\frac{\sqrt{2\sigma^2}T\psi\Gamma(\frac{d+1}{2})}{9.69\Gamma(\frac{d}{2})\epsilon'}\right)^d.$$

□

A.7 Expanded Experiments

Datasets Details

HELOC. The FICO HELOC [36] dataset contains anonymized information about home equity line of credit applications made by homeowners in the US, with a binary response indicating whether or not the applicant has ever been more than 90 days delinquent for a payment. It can be used to train a machine learning model to predict whether the homeowner qualifies for a line of credit or not. The dataset consists of 10459 rows and 40 features, which we have normalized to be between zero and one.

German Credit. The German Credit Dataset [37] comprises 1000 entries, each representing an individual who has taken credit from a bank. These entries are characterized by 20 categorical features, which are used to classify each person as a good or bad credit risk. To prepare the dataset for analysis, we one-hot encoded the data and normalized it such that all features fall between 0 and 1. Additionally, we partitioned the dataset into a training set and a test set, with a 70:30 ratio respectively.

CTG. The CTG dataset [37] consists of 2126 fetal cardiotocograms, which have been evaluated and categorized by experienced obstetricians into three categories: healthy, suspect, and pathological. We

process this dataset based on [29]. The problem was transformed into a binary classification task, where healthy fetuses are distinguished from the other two categories. We divided the dataset into a training set of 1,700 instances and a validation set of 425 instances. Each instance is described by 21 features, which we normalized to have values in $[0, 1]$.

Adult. The Adult dataset [37] is a publicly available dataset in the UCI repository based on 1994 U.S. census data. This dataset is a classification task to successfully predict whether an individual earns more or less than $50K$ per year based on features such as occupation, education, etc. The dataset consists of approximately 48,842 instances, split into a training set of 32,561 instances and a test set of 16,281 instances.

Taiwanese Credit. The Taiwanese dataset [38] consists of 30,000 instances with 24 features that include individuals' financial data, with a binary response indicating their creditworthiness. We use one-hot encoding on the data and normalize it to be in $[0, 1]$. This dataset is processed based on [29]. We partition the data into a training set of 22,500 and a test set of 7,500.

Model Architecture

We first trained a base neural network model for which we aim to find counterfactuals for instances with false negative predictions. The architecture of our base model consisted of two hidden layers, each containing 128 hidden units. We employed the rectified linear unit (ReLU) activation function and Adam optimizer. The model was trained for 50 epochs using a batch size of 32. We employed this model architecture and training setup on all datasets since it yielded a satisfactory level of accuracy on all of them. To evaluate the robustness of these counterfactual examples, we then trained 50 new models and assessed the validity of the counterfactuals under various model change scenarios.

All 50 models had the same architecture and training setup as the base model except for some slight changes which include: Weight Initialization (WI): Retraining new models with the same hyperparameters but different weight initialization by using different random seeds for each model. Leave Out (LO): Retraining new models by randomly removing a small portion (1%) of the training data each time and using different weight initialization.

Implementation Details

The stability measure has the number of samples k and the variance σ^2 as hyperparameters. Our theoretical findings suggest that a higher value of k improves the robustness of the counterfactuals but at the expense of increased cost and computational complexity. In our experiments, we found that a value of $k = 1000$ was sufficient. The value of σ^2 was determined by analyzing the variance of the features in the dataset, and we found that a value of $\sigma^2 = 0.01$ produced good results for the features that lie between $[0, 1]$. These hyperparameters were kept constant across all our experiments and datasets. The baseline technique for generating counterfactuals used in T-rex was the min Cost counterfactual [25]. For other algorithm parameters, a step size of 0.01 was fixed for all datasets and experiments. The maximum number of iterations (max_steps) varied depending on the dataset, with 50 for HELOC, 200 for German Credit, and 100 for CTG, Adult, and Taiwanese. An appropriate value of τ balances the trade-off between validity and cost. We choose a value of τ to guarantee high validity and compare cost and LOF with the baselines. Another method of choosing τ is to use the histogram of $R(\cdot)$ on the training dataset (e.g., see Fig. A.1 for HELOC dataset). To implement ROAR and SNS, we adhered to techniques and algorithm parameters discussed in the original works [28, 29]. In NN and T-Rex:NN implementation, a crucial consideration is determining the appropriate number of

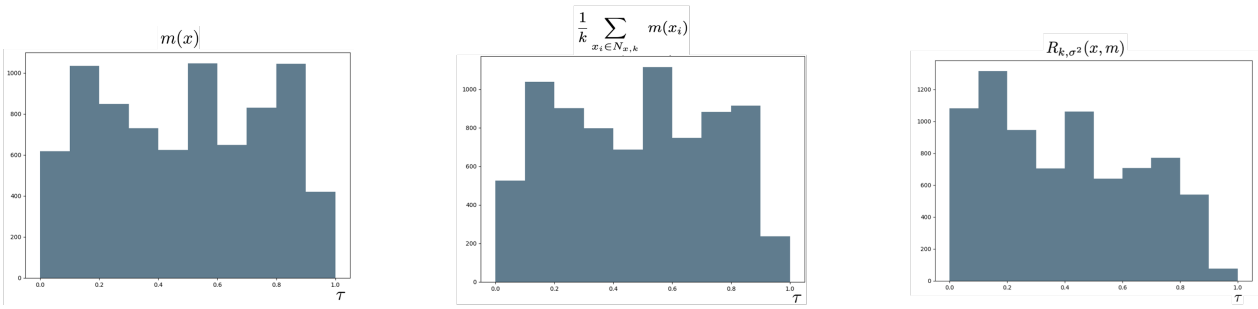


Figure A.1: Histograms on the HELOC dataset to visualize the proposed stability measure.

neighbors K, to search for a robust counterfactual. For larger datasets such as HELOC, Adult, and Taiwanese datasets $K = 1000$, while for the German credit and CTG datasets $K = 100$. Note that if K is too small, a counterfactual robustness test might not be satisfied, and hence T-Rex:NN returns no robust counterfactuals.

Additional Experimental Results

Table A.1: Experimental results for HELOC dataset with standard deviations

HELOC	*l_1 based*				*l_2 based*			
	COST	LOF	WI VAL.	LO VAL.	COST	LOF	WI VAL.	LO VAL.
min Cost	0.40 ± 0.48	0.49 ± 0.93	$38.8 \pm 2.5\%$	$35.2 \pm 2.5\%$	0.11 ± 0.08	0.75 ± 0.61	$13.5 \pm 3.4\%$	$13.5 \pm 3.5\%$
+T-Rex:I	1.02 ± 0.41	0.38 ± 0.95	$98.2 \pm 1.2\%$	$98.1 \pm 1.2\%$	0.29 ± 0.07	0.68 ± 0.64	$98.5 \pm 1.9\%$	$98.2 \pm 2.0\%$
+SNS	1.20 ± 0.46	0.30 ± 0.46	$98.0 \pm 1.2\%$	$97.8 \pm 1.1\%$	0.31 ± 0.08	0.64 ± 0.78	$97.9 \pm 0.8\%$	$97.0 \pm 0.8\%$
ROAR	1.69 ± 1.59	0.41 ± 0.73	$92.6 \pm 3.9\%$	$91.2 \pm 4.3\%$	1.91 ± 2.22	0.43 ± 0.83	$86.3 \pm 0.2\%$	$84.8 \pm 0.3\%$
NN	1.91 ± 2.27	0.80 ± 2.27	$51.1 \pm 12.6\%$	$50.3 \pm 12.4\%$	0.56 ± 0.59	0.80 ± 2.27	$51.1 \pm 12.6\%$	$50.3 \pm 12.4\%$
T-Rex:NN	2.50 ± 1.83	0.92 ± 0.56	$84.0 \pm 0.8\%$	$84.0 \pm 0.8\%$	0.77 ± 0.61	0.92 ± 0.56	$84.0 \pm 0.8\%$	$84.0 \pm 0.8\%$

Table A.2: Experimental results for German Credit dataset with standard deviations.

GERMAN	*l_1 based*				*l_2 based*			
	COST	LOF	WI VAL.	LO VAL.	COST	LOF	WI VAL.	LO VAL.
min Cost	1.42 ± 4.16	0.77 ± 0.65	$58.8 \pm 7.9\%$	$56.7 \pm 7.3\%$	0.48 ± 1.38	0.81 ± 0.71	$26.6 \pm 15\%$	$26.6 \pm 14\%$
+T-Rex:I	4.81 ± 3.93	0.72 ± 0.73	$98.0 \pm 4.6\%$	$96.5 \pm 3.8\%$	1.20 ± 1.37	0.75 ± 0.71	$99.2 \pm 0.9\%$	$98.7 \pm 0.9\%$
+SNS	5.71 ± 4.10	0.67 ± 0.39	$97.5 \pm 0.3\%$	$98.1 \pm 0.2\%$	1.44 ± 1.38	0.68 ± 0.73	$99.9 \pm 0.0\%$	$98.9 \pm 0.1\%$
ROAR	7.63 ± 4.00	0.54 ± 0.67	$96.3 \pm 0.2\%$	$92.3 \pm 0.3\%$	6.81 ± 1.00	0.58 ± 0.98	$87.8 \pm 1.4\%$	$85.2 \pm 1.6\%$
NN	7.05 ± 3.96	1.00 ± 0.00	$95.3 \pm 0.9\%$	$95.4 \pm 1.0\%$	2.50 ± 1.20	1.00 ± 0.00	95.3 ± 0.9	$95.3 \pm 1.0\%$
T-Rex:NN	10.13 ± 4.10	1.00 ± 0.00	$100 \pm 0.0\%$	$100 \pm 0.0\%$	3.04 ± 1.45	1.00 ± 0.00	$100 \pm 0.0\%$	$100 \pm 0.0\%$

Table A.3: Experimental results for CTG dataset with standard deviations.

CTG	l_1 *based*				l_2 *based*			
	COST	LOF	WI VAL.	LO VAL.	COST	LOF	WI VAL.	LO VAL.
min Cost	0.21 ± 0.18	0.94 ± 0.91	$74.6 \pm 0.1\%$	$70.2 \pm 0.2\%$	0.08 ± 0.04	1.00 ± 0.00	$19.7 \pm 30\%$	$14.1 \pm 31\%$
+T-Rex:I	1.11 ± 0.11	0.83 ± 0.91	$100 \pm 0.0\%$	$98.8 \pm 0.1\%$	0.42 ± 0.04	0.94 ± 0.63	$100 \pm 0.0\%$	$99.7 \pm 0.1\%$
+SNS	3.34 ± 0.18	-1.00 ± 0.0	$100 \pm 0.0\%$	$98.2 \pm 0.1\%$	1.07 ± 0.04	-1.00 ± 0.0	$100 \pm 0.0\%$	$99.3 \pm 0.1\%$
ROAR	3.68 ± 3.48	0.64 ± 0.78	$98.7 \pm 0.5\%$	$96.4 \pm 0.3\%$	1.35 ± 2.01	0.59 ± 0.90	$98.8 \pm 0.0\%$	$97.2 \pm 0.0\%$
NN	0.39 ± 0.23	1.00 ± 0.00	$70.5 \pm 0.2\%$	$67.5 \pm 0.1\%$	0.15 ± 0.07	1.00 ± 0.00	$70.5 \pm 0.2\%$	$67.5 \pm 0.1\%$
T-Rex:NN	2.22 ± 0.12	-0.33 ± 0.67	$100 \pm 0.0\%$	$100 \pm 0.0\%$	1.00 ± 0.80	-0.33 ± 0.67	$100 \pm 0.0\%$	$100 \pm 0.0\%$

Table A.4: Experimental results for Taiwanese Credit dataset with standard deviations.

TAIWANESE	l_1 *based*				l_2 *based*			
	COST	LOF	WI VAL.	LO VAL.	COST	LOF	WI VAL.	LO VAL.
min Cost	3.95 ± 3.42	-0.37 ± 0.92	$38.4 \pm 6.0\%$	$38.4 \pm 5.9\%$	2.84 ± 1.16	-0.68 ± 0.72	$21.1 \pm 2.1\%$	$20.0 \pm 2.2\%$
+T-Rex:I	6.34 ± 3.26	0.48 ± 0.67	$96.9 \pm 0.7\%$	$96.2 \pm 0.7\%$	3.06 ± 1.11	0.40 ± 0.31	$96.8 \pm 2.1\%$	$96.4 \pm 1.9\%$
+SNS	6.51 ± 3.37	0.39 ± 0.39	$97.2 \pm 1.3\%$	$96.9 \pm 1.4\%$	3.10 ± 1.16	0.43 ± 0.62	$96.7 \pm 3.1\%$	$96.1 \pm 2.7\%$
NN	5.80 ± 3.82	0.84 ± 0.76	$53.5 \pm 9.9\%$	$51.6 \pm 8.9\%$	2.18 ± 1.27	0.84 ± 0.76	$53.5 \pm 9.9\%$	$51.6 \pm 8.9\%$
T-Rex:NN	6.89 ± 4.51	0.86 ± 0.61	$98.8 \pm 0.9\%$	$98.4 \pm 0.8\%$	3.54 ± 1.48	0.86 ± 0.61	$98.8 \pm 0.9\%$	$98.4 \pm 0.8\%$

Table A.5: Experimental results for Adult dataset with standard deviations.

ADULT	l_1 *based*				l_2 *based*			
	COST	LOF	WI VAL.	LO VAL.	COST	LOF	WI VAL.	LO VAL.
min Cost	0.12 ± 0.54	0.09 ± 0.99	$80.8 \pm 5.0\%$	$78.2 \pm 4.7\%$	0.18 ± 0.65	0.09 ± 0.99	$50.0 \pm 8.0\%$	$49.2 \pm 8.2\%$
+T-Rex:I	0.51 ± 0.53	0.08 ± 0.99	$98.4 \pm 1.6\%$	$98.1 \pm 1.3\%$	0.24 ± 0.64	0.09 ± 0.99	$98.2 \pm 0.0\%$	$98.2 \pm 0.0\%$
+SNS	0.92 ± 0.58	-0.2 ± 0.21	$98.6 \pm 0.0\%$	$97.9 \pm 0.1\%$	0.37 ± 0.62	-0.22 ± 0.97	$97.9 \pm 0.0\%$	$97.8 \pm 0.0\%$
NN	2.16 ± 1.43	0.91 ± 0.03	$81.0 \pm 2.8\%$	$81.0 \pm 2.7\%$	1.21 ± 0.65	0.91 ± 0.03	$81.0 \pm 2.8\%$	$81.0 \pm 2.7\%$
T-Rex:NN	3.25 ± 1.6	0.85 ± 0.02	$99.2 \pm 0.0\%$	$99.0 \pm 0.0\%$	1.59 ± 0.57	0.85 ± 0.02	$99.2 \pm 0.0\%$	$99.0 \pm 0.0\%$

Table A.6: Ablation study on German Credit Dataset.

		l_1 based			l_2 based		
τ	Measure	COST	LOF	WI VAL.(%)	COST	LOF	WI VAL.(%)
0.5	$r(x,m)$	1.12	0.78	57.0	0.54	0.81	32.6
	$r_{k,\sigma^2}(x,m)$	2.06	0.81	57.4	0.57	0.80	37.3
	$R_{k,\sigma^2}(x,m)$	2.33	0.80	66.9	0.65	0.80	52.0
0.6	$r(x,m)$	1.48	0.77	61.7	0.56	0.80	38.9
	$r_{k,\sigma^2}(x,m)$	2.11	0.77	62.5	0.58	0.79	43.0
	$R_{k,\sigma^2}(x,m)$	2.40	0.76	72.9	0.68	0.81	61.7
0.7	$r(x,m)$	1.73	0.71	61.3	0.63	0.87	42.8
	$r_{k,\sigma^2}(x,m)$	2.61	0.75	63.7	0.67	0.87	51.8
	$R_{k,\sigma^2}(x,m)$	2.90	0.76	72.6	0.76	0.88	67.0
0.8	$r(x,m)$	1.69	0.76	74.4	0.68	0.85	61.0
	$r_{k,\sigma^2}(x,m)$	2.50	0.79	79.0	0.74	0.85	73.3
	$R_{k,\sigma^2}(x,m)$	2.77	0.77	86.0	0.82	0.84	84.2
0.9	$r(x,m)$	2.24	0.73	79.5	0.76	0.81	71.0
	$r_{k,\sigma^2}(x,m)$	3.01	0.74	84.8	0.83	0.79	82.7
	$R_{k,\sigma^2}(x,m)$	3.30	0.74	89.6	0.91	0.78	89.0

Appendix B

B.1 Background on Stability Measure

How stability differs from existing robustness measures: Our focus on model multiplicity distinguishes this work from traditional robustness measures, which address different aspects of model behavior such as out-of-distribution (OOD) generalization, stability under natural perturbations, and uncertainty estimation [103]. OOD generalization typically evaluates how well a model performs on data that differs from the training distribution (e.g., classifying objects seen from novel viewpoints or in cluttered settings). This is often quantified using test datasets with altered conditions or domain shifts, and methods like domain adaptation are employed to enhance robustness. Stability under natural perturbations assesses the sensitivity of predictions and predicted probabilities to small, random changes in the input, such as Gaussian noise or image transformations. Uncertainty estimation, on the other hand, focuses

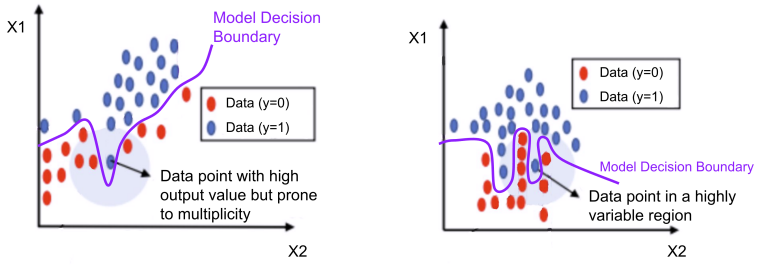


Figure B.1: Additional motivation for our local stability measure. Our measure relies on both local variability and mean confidence as they capture synergistic aspects of prediction robustness.

on calibrating the predicted probabilities to reflect true likelihoods, often using measures like Expected Calibration Error or entropy-based metrics to evaluate how well the model quantifies confidence in its predictions. While these methods provide valuable insights into different facets of robustness, their goals differ significantly from ours.

[104] is more closely related to our approach, as it quantifies robustness by measuring the fraction of stable predictions within a local neighborhood. While both approaches leverage the neighborhood around a data point, the objectives diverge: [104] focuses on quantifying the probability of stable predictions against perturbations to evaluate robustness to noise. In contrast, our measure aims to capture the consistency of predictions (multiplicity) among competing models within the Rashomon set. Additionally, our stability measure's unique mean-variance nature further distinguishes it (see Figure B.1). Unlike existing metrics, it not only accounts for the average prediction within a neighborhood but also penalizes the variability in predictions. Moreover, we provide theoretical guarantees on the consistency of predictions with high stability scores over a broad range of equally-well performing models. Recent work has also explored consistency in LLMs across repeated inference runs or under slight semantic perturbations to the input [122, 124, 141, 316].

B.2 Proof of Theorem 3.1

Theorem 3.1 (Probabilistic Guarantee)**.** *Given* $\mathbf{x}$*, a target model* $\bar{f}$ *and stability measure* $S_{k,\sigma}(\mathbf{x}, F)$*. Under Assumption 3.1, and for all* $\epsilon > \epsilon'$*, where* $\epsilon' = 2(\alpha + t\sigma) + \mathcal{O}(L\sigma^2)$*. We have:*

$$\Pr\left(\bar{f}(\mathbf{x}) \geq S_{k,\sigma}(\mathbf{x}, F) - \epsilon\right) \geq 1 - \exp\left(\frac{-k\epsilon^2}{32}\right) \tag{3.5}$$

Proof. To prove Theorem 3.1, we begin with Lemma B.1.

Assume the fine-tuned models F belong to a discrete class of random variables. A specific model realization is represented as f^i for $i = 1, 2, \ldots, |\mathcal{F}_\delta|$, with the complete set denoted by $\mathcal{F} = \{f^1, f^2, \ldots, f^{|\mathcal{F}|}\}$. Each model f^i is selected with probability p_i, where $\sum_{i=1}^{|\mathcal{F}_\delta|} p_i = 1$.

Lemma B.1. *Given* $Z_i = F(X_i) - \bar{f}(X_i) - |\bar{f}(X_i) - \bar{f}(\mathbf{x})| + |F(X_i) - F(\mathbf{x})|$ *and* $Z = \frac{1}{k}\sum_{i=1}^{k} Z_i$, *under Assumption 3.1, for any* $\tilde{\epsilon} > \epsilon' > 0$, *we have:*

$$\Pr(Z \geq \epsilon' + \tilde{\epsilon}) \leq \exp\left(\frac{-k(\tilde{\epsilon} + \epsilon')^2}{32}\right). \tag{B.1}$$

Lemma B.2 (Hoeffding's Inequality)**.** *For a given random variable* X_i *such that* $X_i \in [a, b]$ *almost surely, and for any* $\varepsilon > 0$,

$$\Pr\left(\left|\frac{1}{k}\sum_{i=1}^{k} X_i - \mathbb{E}(X_i)\right| > \varepsilon\right) \leq 2\exp\left(-\frac{2k\varepsilon^2}{(b-a)^2}\right). \tag{B.2}$$

See [317] for detailed proof of Hoeffding's Inequality.

Since $\bar{f}(\cdot), F(\cdot) \in [0, 1]$, we have $Z_i \in [-2, 2]$. Hence, from Lemma B.2, we have:

$$\Pr\left(|Z - \mathbb{E}[Z|F = f]| \geq \tilde{\epsilon} \mid F = f\right) \leq 2\exp\left(-\frac{k\tilde{\epsilon}^2}{8}\right) \tag{B.3}$$

Given $|\mathbb{E}[Z|F = f]| < \epsilon'$, we have $-\epsilon' < \mathbb{E}[Z|F = f] < \epsilon'\ \forall f$ (see Lemma B.3).

Now observe that:

$$\Pr(Z \geq \epsilon' + \tilde{\epsilon}|F = f) \overset{\text{(a)}}{\leq} \Pr(Z \geq \mathbb{E}[Z|F{=}f] + \tilde{\epsilon}|F{=}f) \leq \exp\left(\frac{-k\tilde{\epsilon}^2}{8}\right). \tag{B.4}$$

Here, (a) holds since $\mathbb{E}[Z|F = f] < \epsilon'$. The event on the left is a subset of that on the right.

Therefore, the probability of the event $\{Z \geq \epsilon' + \tilde{\epsilon}\}$ occurring cannot be more than the probability of the event $\{Z \geq \mathbb{E}[Z|F = f] + \tilde{\epsilon}\}$ occurring.

$$\Pr(Z \geq \epsilon' + \tilde{\epsilon}) \overset{\text{(b)}}{=} \sum_i \Pr(Z \geq \epsilon' + \tilde{\epsilon} | F = f^i) \Pr(F = f^i) \tag{B.5}$$

$$\overset{\text{(c)}}{\leq} \exp\left(\frac{-k\tilde{\epsilon}^2}{8}\right) \sum_i \Pr(F = f^i) \tag{B.6}$$

$$= \exp\left(\frac{-k\tilde{\epsilon}^2}{8}\right) \tag{B.7}$$

$$\overset{\text{(d)}}{\leq} \exp\left(\frac{-k(\tilde{\epsilon} + \epsilon')^2}{32}\right) \tag{B.8}$$

Here, (b) holds from the law of total probability. Next, (c) follows from (B.4). Finally, (d) holds from using the inequality $4\tilde{\epsilon}^2 > (\tilde{\epsilon} + \epsilon')^2$ which holds for $\tilde{\epsilon} > \epsilon' > 0$. Setting $\epsilon = \tilde{\epsilon} + \epsilon'$.

We have:

$$\Pr\left(\frac{1}{k}\sum_{i=1}^{k} \bar{f}(X_i) \geq \frac{1}{k}\sum_{i=1}^{k} \left(F(X_i) - |F(X_i) - F(\mathbf{x})| + |\bar{f}(X_i) - \bar{f}(\mathbf{x})|\right) - \epsilon\right) \geq 1 - \exp\left(\frac{-k\epsilon^2}{32}\right). \tag{B.9}$$

Observe that $\bar{f}(\mathbf{x}) \geq \bar{f}(\mathbf{x}_i) - |\bar{f}(\mathbf{x}_i) - \bar{f}(\mathbf{x})|$. This applies directly from the reverse triangle inequality, i.e., for any real numbers a and b, we have:$|a| \geq |b| - |a - b|$. Hence,

$$\bar{f}(\mathbf{x}) \geq \frac{1}{k}\sum_{i=1}^{k} (\bar{f}(X_i) - |\bar{f}(X_i) - \bar{f}(\mathbf{x})|) \tag{B.10}$$

Therefore, plugging (B.10) into (B.9), we have:

$$\Pr\left(\bar{f}(\mathbf{x}) \geq \frac{1}{k}\sum_{i=1}^{k}(F(X_i) - |F(X_i) - F(\mathbf{x})| + |\bar{f}(X_i) - \bar{f}(\mathbf{x})| - |\bar{f}(X_i) - \bar{f}(\mathbf{x})| - \epsilon)\right) \tag{B.11}$$

$$= \Pr\left(\bar{f}(\mathbf{x}) \geq \frac{1}{k}\sum_{i=1}^{k}(F(X_i) - |F(X_i) - F(\mathbf{x})|) - \epsilon\right) \geq 1 - \exp\left(\frac{-k\epsilon^2}{32}\right). \tag{B.12}$$

Given $S_{k,\sigma}(\mathbf{x}, F) = \frac{1}{k}\sum_{i=1}^{k}(F(X_i) - |F(\mathbf{x}) - F(X_i)|)$, we have:

$$\Pr\left(\bar{f}(\mathbf{x}) \geq S_{k,\sigma}(\mathbf{x}, F) - \epsilon\right) \geq 1 - \exp\left(\frac{-k\epsilon^2}{32}\right). \tag{B.13}$$

Using Lemma B.3, we show that $\mathbb{E}[Z|F = f] \leq \epsilon'$ which completes the proof.

Lemma B.3. *Let $F, \bar{f}$ be twice differentiable functions with Hessians bounded by L, i.e., $\|\nabla^2 F(\mathbf{x})\|, \|\nabla^2 \bar{f}(\mathbf{x})\| \leq L$ for all $\mathbf{x}$. Let $\mathbb{E}_{X_i}[\|F(X_i) - \bar{f}(X_i)\||F = f] \leq \alpha$ and $\mathbb{E}_{X_i}[\|\nabla(F - \bar{f})(X_i)\||F = f] \leq t$, where X_i is drawn uniformly from distribution over the hypersphere $B(\mathbf{x}, \sigma)$. If $Z_i = \big[F(X_i) - \bar{f}(X_i)\big] - \big|\bar{f}(X_i) - \bar{f}(\mathbf{x})\big| + \big|F(X_i) - F(\mathbf{x})\big|$, then,*

$$\mathbb{E}[Z_i|F = f] \leq \alpha + t\sigma + \underbrace{\mathcal{O}(L\sigma^2)}_{\textit{Hessian Error}}, \tag{B.14}$$

Proof. We first bound $\mathbb{E}[F(X_i) - \bar{f}(X_i)]$:

Since $\mathbb{E}[|F(X_i) - \bar{f}(X_i)|] \leq \alpha$ for all it follows that $\mathbb{E}[F(X_i) - \bar{f}(X_i)] \leq \alpha$.

Next, we bound $\mathbb{E}\big[|F(X_i) - F(\mathbf{x})| - |\bar{f}(X_i) - \bar{f}(\mathbf{x})|\big]$:

Expand $F(X_i)$ and $\bar{f}(X_i)$ around $\mathbf{x}$ using Taylor's expansion:

$$
\begin{aligned}
F(\mathbf{x}) &= F(X_i) + \nabla F(X_i)^T(\mathbf{x} - X_i) + \frac{1}{2}(\mathbf{x} - X_i)^T \nabla^2 F(\xi_F)(\mathbf{x} - X_i), \\
\bar{f}(\mathbf{x}) &= \bar{f}(X_i) + \nabla \bar{f}(X_i)^T(\mathbf{x} - X_i) + \frac{1}{2}(\mathbf{x} - X_i)^T \nabla^2 \bar{f}(\xi_{\bar{f}})(\mathbf{x} - X_i),
\end{aligned} \tag{B.15}
$$

for some $\xi_F, \xi_{\bar{f}} \in B(\mathbf{x}, \sigma)$. The absolute differences become:

$$
\begin{aligned}
|F(X_i) - F(\mathbf{x})| &\leq \left|\nabla F(X_i)^T(\mathbf{x} - X_i)\right| + \frac{1}{2}L\|X_i - \mathbf{x}\|^2, \\
|\bar{f}(X_i) - \bar{f}(\mathbf{x})| &\geq \left|\nabla \bar{f}(X_i)^T(\mathbf{x} - X_i)\right| - \frac{1}{2}L\|X_i - \mathbf{x}\|^2.
\end{aligned} \tag{B.16}
$$

Subtracting these:

$$
|F(X_i) - F(\mathbf{x})| - |\bar{f}(X_i) - \bar{f}(\mathbf{x})| \leq \left|\nabla F(X_i)^T(\mathbf{x} - X_i)\right| - \left|\nabla \bar{f}(X_i)^T(\mathbf{x} - X_i)\right| + L\|X_i - \mathbf{x}\|^2. \tag{B.17}
$$

Using the reverse triangle inequality $|a| - |b| \leq |a - b|$:

$$
|F(X_i) - F(\mathbf{x})| - |\bar{f}(X_i) - \bar{f}(\mathbf{x})| \leq \left|\nabla (F - \bar{f})(X_i)^T(\mathbf{x} - X_i)\right| + L\|X_i - \mathbf{x}\|^2. \tag{B.18}
$$

Taking Expectation:

$$
\mathbb{E}\left[\left|\nabla (F - \bar{f})(X_i)^T(\mathbf{x} - X_i)\right|\right] \leq \mathbb{E}[\|\nabla (F - \bar{f})(X_i)\|\|X_i - \mathbf{x}\|] \leq \mathbb{E}[\|\nabla (F - \bar{f})(X_i)\|\sigma \leq t\sigma \tag{B.19}
$$

Hessian Term: $\mathbb{E}\left[L\|X_i - \mathbf{x}\|^2\right] = L \cdot \mathbb{E}[\|X_i - \mathbf{x}\|^2] = \mathcal{O}(L\sigma^2)$.

Combining the inequalities, we have:

$$
\mathbb{E}[Z_i | F = f] \leq \alpha + t\sigma + \mathcal{O}(L\sigma^2). \tag{B.20}
$$

□

□

B.3 Expanded Experiments

Definition B.1 (Spearman Correlation)**.** *Spearman*(X, Y) *measures the strength and direction of a monotonic relationship between two variables. It is the Pearson correlation coefficient of their ranked values. Given* n *pairs* (X_i, Y_i)*, it is computed as:*

$$\textit{Spearman}(X, Y) = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} = \frac{\textit{cov}(\textit{rank}(X), \textit{rank}(Y))}{\sigma_{\textit{rank}(X)} \sigma_{\textit{rank}(Y)}},$$

where d_i *is the difference between the ranks of* X_i *and* Y_i*. The value ranges from* -1 *(perfect negative monotonicity) to* 1 *(perfect positive monotonicity), with* 0 *indicating no monotonic relationship.*

Dataset Details

This section provides detailed descriptions of the datasets used in our experiments.

Adult dataset [37], also known as the "Census Income" dataset, is used for predicting whether an individual earns more than $50,000 annually based on various demographic attributes. It consists of 48,842 instances with 14 attributes, such as age, work class, education, occupation, capital gain, capital loss, hours per week, etc. The dataset is commonly used in classification tasks.

German Credit dataset [37] is used for credit risk evaluation. It consists of $1,000$ instances with 20 attributes, which include personal information, credit history, and loan attributes. The target variable indicates whether the credit is good or bad. This dataset is often used for binary classification problems

and helps in understanding the factors affecting creditworthiness.

Diabetes dataset [80] is used for predicting the onset of diabetes based on diagnostic measurements. It contains 768 instances with 8 attributes, including the number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. The target variable indicates whether the individual has diabetes. The dataset is commonly used in classification tasks.

Bank dataset [107] is used for predicting whether a client will subscribe to a term deposit based on data from direct marketing campaigns of a Portuguese bank. It includes 45,211 instances in the training set and 18 attributes, such as age, job type, marital status, education, credit balance, housing loan status, and contact details from the marketing campaigns. The target variable indicates whether the client subscribed to the term deposit. This dataset is commonly used in binary classification tasks.

Heart dataset [318] contains data from four different hospitals. It includes 918 patients, each represented by 11 clinical variables, with the task being a binary classification of coronary artery disease. Among the patients, 508 are labeled positive for the condition.

Car dataset [319] contains entries describing various cars characterized by six attributes. The task is a classification problem aimed at evaluating the state of each car. The dataset comprises 1,728 examples.

Hospital Remission dataset [114] contains 99,493 inpatient encounters from 130 U.S. hospitals (1999 – 2008) that involve diabetic patients. The dataset provides 50+ attributes aimed at predicting whether a patient will be readmitted within 30 days after discharge. It is commonly used for binary classification tasks assessing post-discharge risk and care quality.

Experimental Setup

Our experiments were carried out using the BigScience T0 and Google Flan T5 models fine-tuned on several datasets. The number of shots was set to 64,128, and 512 for each dataset. To evaluate multiplicity and local stability, we fine-tuned 40 models with different random seeds for each dataset and recorded their predictions. The training process involved setting the batch size to 2 for smaller training sizes and 8 for larger sizes. The learning rate was set to 0.003. For each dataset, we determined the number of training steps adaptively based on the number of shots, ensuring sufficient iterations for model convergence. Specifically, the training steps were calculated as $20 \times (\text{number of shots}/\text{batch size})$. All experiments were performed on 2 NVIDIA RTX A4500 and 4 NVIDIA RTX 6000 GPUs. To ensure reproducibility and robustness of the results, different random seeds (i.e., 2, 4, 8, etc) were used for each fine-tuning iteration. For fine-tuning with LoRA we use a rank of 4. Given the infeasibility of computing the exact size of $|\mathcal{F}_\delta|$ due to its potentially vast model space, we employ an expensive sampling approach, i.e., fine-tuning with various seeds. We select a finite number of models from $\mathcal{F}_\delta$ for practical evaluation, allowing us to evaluate the multiplicity metrics. It is very computationally expensive to fine-tune several models to evaluate multiplicity. This motivates the need for a measure to quantify stability given one model.

Expanded Results

This section presents a broader set of experimental results. We evaluate multiplicity across several datasets using both the BigScience T0 model fine-tuned with LoRA (see Table B.1) and the FLAN-T5 model fine-tuned using the Tfew recipe (see Table B.2). Additionally, Figures B.2, B.3, and B.4 visualize the evaluated multiplicity versus the stability measure for Bank, Diabetes, and Ger-

Table B.1: Multiplicity Evaluation Metrics for Different Datasets and Number of Shots. Evaluated on 40 fine-tuned **BigScience T0** models on **LoRA** using different random seeds. Multiplicity observed in predictions across different fine-tuned model, even when models exhibit similar accuracy (in this setting $\delta = 0.02$).

Dataset	No. Shots	Multiplicity Evaluation Metrics (BigScience T0)					
		Arbitrariness	Discrepancy	Avg. Pairwise Disagreement	Avg. Pred. Variance	Avg. Pred. Range	Avg. Model Accuracy
Adult	64	11%	6%	9%	0.01	0.11	83%
	128	10%	9%	6%	0.01	0.10	84%
	512	11%	3%	10%	0.01	0.12	85%
German	64	19%	10%	6%	0.04	0.40	70%
	128	17%	11%	6%	0.01	0.16	71%
	512	21%	14%	8%	0.03	0.26	72%
Diabetes	64	20%	13%	11%	0.04	0.21	70%
	128	16%	14%	11%	0.08	0.14	73%
	512	19%	13%	11%	0.04	0.17	76%
Bank	64	13%	9%	7%	0.01	0.28	66%
	128	14%	9%	7%	0.03	0.21	73%
	512	14%	8%	7%	0.03	0.22	78%

man Credit datasets respectively. We also report the correlation between the stability measure and various multiplicity evaluation metrics for BigScience T0 model fine-tuned using Tfew recipe (see Table B.3) and FLAN-T5 model fine-tuned using Tfew recipe (see Table B.4).

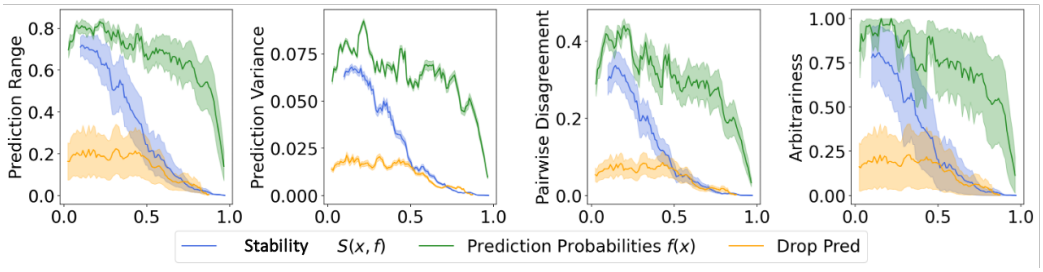


Figure B.2: Evaluated multiplicity (assessed on 40 retrained models) versus our stability measure (evaluated on one model) for the **512-shot** setting on the **Bank dataset**. The plots demonstrate that high stability values correspond to low multiplicity across various multiplicity evaluation metrics. Predictive probabilities and Drop-Out not providing any providing any useful insight into multiplicity.

Table B.2: Evaluated Multiplicity for Different Datasets and Number of Shots. Evaluated on 40 fine-tuned **FLAN-T5** models using **Tfew** recipe with different random seeds. Multiplicity observed in predictions across different fine-tuned models, even when models exhibit similar accuracy (in this setting $\delta = 0.02$). The accuracy of FLAN T5 model on the dataset is less than the BigScience T0 model observed in Table 3.1.

Dataset	No. Shots	Multiplicity Evaluation Metrics (FLAN-T5)					
		Arbitrariness	**Discrepancy**	**Avg. Pairwise Disagreement**	**Avg. Pred. Variance**	**Avg. Pred. Range**	**Avg. Model Accuracy**
Adult	64	13.96%	6.93%	5.05%	0.010	0.139	74.25%
	128	8.81%	3.84%	3.39%	0.008	0.091	77.50%
	512	12.02%	5.71%	4.49%	0.012	0.123	79.17%
German	64	18.50%	11.00%	6.19%	0.015	0.194	64.85%
	128	30.00%	13.50%	10.47%	0.031	0.287	69.25%
	512	35.50%	16.50%	12.88%	0.041	0.362	69.40%
Diabetes	64	15.58%	7.79%	6.23%	0.016	0.170	68.18%
	128	11.69%	5.84%	4.81%	0.012	0.129	59.29%
	512	21.43%	9.74%	7.37%	0.022	0.207	69.55%
Bank	64	12.86%	7.46%	4.69%	0.003	0.125	66.96%
	128	17.95%	6.90%	6.59%	0.006	0.165	65.94%
	512	17.17%	6.61%	6.24%	0.017	0.173	79.40%

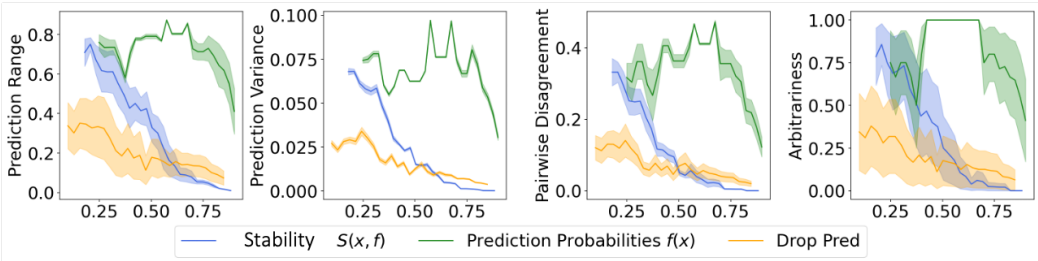


Figure B.3: Evaluated multiplicity (assessed on 40 retrained models) versus our stability measure (evaluated on one model) for the **512-shot** setting on the **Diabetes dataset**. The plots demonstrate that high stability values correspond to low multiplicity across various multiplicity evaluation metrics. Predictive probabilities not providing any providing any useful insight about multiplicity. The drop-out method performs better than predictive probabilities but still worse than stability.

Expanded Ablations

We include the complete ablation results referenced in the main paper. Table B.6 presents the results of varying the sample size k. Table B.7 and Figures B.5 examines the effect of different values of σ on the stability measure. Additionally, Table B.8 compares the stability measure with the drop-

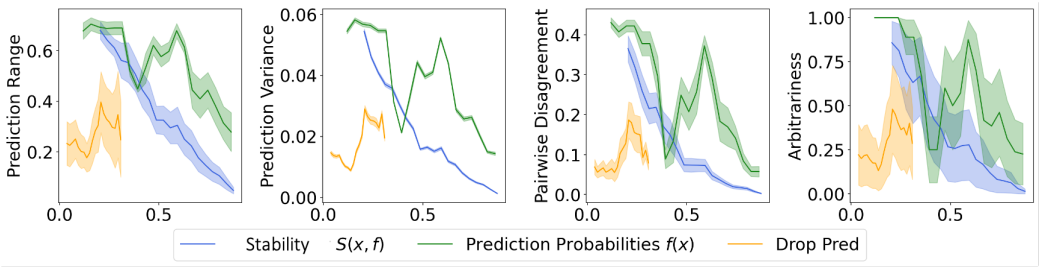


Figure B.4: Evaluated multiplicity (assessed on 40 retrained models) versus our stability measure (evaluated on one model) for the **512-shot** setting on the **German Credit dataset**. The plots demonstrate that high stability values correspond to low multiplicity across various multiplicity evaluation metrics. In this setting Prediction probability is performing competitively. But generally stability measure provides better insight into the multiplicity of predictions compared to the predicted probabilities. The drop-out method is performing significantly worse than the other two measures.

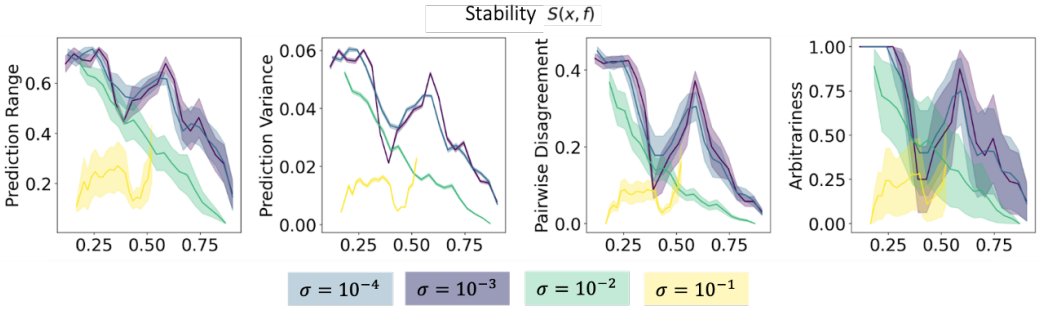


Figure B.5: Ablation study on different σ values: The chosen value of $\sigma = 0.01$ yields the best performance across all evaluation metrics. Smaller values of σ (e.g., $\sigma = 10^{-4}$) result in perturbations that are too close to the original data points, leading to similar outcomes as prediction probability alone, as the sampled points are nearly identical. On the other hand, larger values (e.g., $\sigma = 10^{-2}$) produce overly noisy perturbations, rendering the results uninformative.

out method with different drop-out rates p. Finally, Table B.5 compares the stability measure with variability-based alternatives. Table B.9 compares performance under LoRA, Prompt Tuning, and Prefix Tuning on the Bank dataset.

Table B.3: This table reports the Spearman correlation between the stability measure, predicted probabilities, and the drop-out method with various multiplicity evaluation metrics for different numbers of shots on several datasets (**BigScience T0 fine-tuned using Tfew recipe**). In most cases, the stability measure $S_{k,\sigma}(\mathbf{x}, f)$ shows a higher correlation with these multiplicity measures compared to predicted probabilities and drop-out, indicating that the stability measure $S_{k,\sigma}(\mathbf{x}, f)$ better informs about the multiplicity than the other measures do. The dropout method performing better than naive predicted probability.

Dataset	Number of Shots	Measure	Arbitrariness	Pairwise Disagreement	Prediction Variance	Prediction Range
Adult	64	Pred. Prob.	0.67	0.66	0.50	0.62
		Drop-Out	0.83	0.78	0.81	0.87
		Stability	**0.95**	**0.90**	**0.91**	**0.89**
	128	Pred. Prob.	0.67	0.62	0.30	0.54
		Drop-Out	0.74	0.83	0.69	0.81
		Stability	**0.80**	**0.96**	**0.84**	**0.91**
	512	Pred. Prob.	0.70	0.69	0.56	0.72
		Drop-Out	0.78	0.78	0.88	0.88
		Stability	**0.90**	**0.86**	**0.93**	**0.92**
German Credit	64	Pred. Prob.	**0.99**	**0.99**	0.80	0.79
		Drop-Out	0.73	0.71	0.82	0.76
		Stability	0.95	0.95	**0.98**	**0.84**
	128	Pred. Prob.	**0.57**	**0.57**	0.86	0.86
		Drop-Out	0.50	0.56	0.74	0.84
		Stability	0.54	0.54	**0.87**	**0.87**
	512	Pred. Prob.	0.54	0.56	0.83	0.82
		Drop-Out	**0.69**	**0.67**	0.72	0.65
		Stability	0.59	0.60	**0.87**	**0.86**
Diabetes	64	Pred. Prob.	0.03	0.38	0.04	0.08
		Drop-Out	0.30	0.19	**0.54**	**0.46**
		Stability	**0.45**	**0.51**	0.31	0.23
	128	Pred. Prob.	0.88	0.93	**0.93**	**0.95**
		Drop-Out	0.89	0.92	0.92	0.94
		Stability	**0.92**	**0.95**	**0.93**	**0.95**
	512	Pred. Prob.	0.21	0.23	0.24	0.30
		Drop-Out	0.74	0.83	**0.75**	**0.74**
		Stability	**0.80**	**0.89**	0.74	0.68
Bank	64	Pred. Prob.	0.70	0.69	0.56	0.74
		Drop-Out	0.79	0.77	0.77	**0.80**
		Stability	**0.83**	**0.78**	**0.81**	**0.80**
	128	Pred. Prob.	0.54	0.57	0.73	0.62
		Drop-Out	0.62	0.70	0.75	0.51
		Stability	**0.79**	**0.84**	**0.87**	**0.86**
	512	Pred. Prob.	0.71	0.68	0.81	0.76
		Drop-Out	0.90	0.89	0.87	0.84
		Stability	**0.91**	**0.92**	**0.91**	**0.87**
Heart	64	Pred. Prob.	0.70	0.21	0.30	0.69
		Drop-Out	0.56	0.48	0.54	0.56
		Stability	**0.98**	**0.86**	**0.98**	**0.98**
	128	Pred. Prob.	0.61	0.46	0.50	0.26
		Drop-Out	0.64	0.76	0.74	0.83
		Stability	**0.89**	**0.90**	**0.97**	**0.87**
	512	Pred. Prob.	0.80	0.65	0.48	0.35
		Drop-Out	**0.94**	0.90	**0.90**	0.94
		Stability	0.89	**0.95**	0.86	**0.95**
Car	64	Pred. Prob.	0.83	**0.83**	0.40	0.83
		Drop-Out	**0.85**	**0.83**	**0.96**	**0.97**
		Stability	0.76	0.69	0.86	0.75
	128	Pred. Prob.	0.56	0.26	0.29	0.01
		Drop-Out	0.63	0.66	0.57	0.52
		Stability	**0.97**	**0.91**	**0.93**	**0.94**
	512	Pred. Prob.	0.91	0.94	0.72	0.86
		Drop-Out	**0.98**	**0.96**	**0.95**	**0.93**
		Stability	0.68	0.59	0.56	0.67

Table B.4: Spearman correlation between the predicted probabilities, drop-out method, and the stability measure with various multiplicity evaluation metrics for different numbers of shots on several datasets (**Flan T5 model fine-tuned using Tfew recipe**). In most cases, the stability measure shows a higher correlation with these multiplicity measures compared to predicted probabilities and drop-out, indicating that the stability measure better informs about the multiplicity than the other measures do. The dropout method performs competitively in some cases.

Dataset	Number of Shots	Measure	Arbitrariness	Pairwise Disagreement	Prediction Variance	Prediction Range
Adult	64	Pred. Prob.	0.62	0.67	**0.72**	0.56
		Drop-Out	0.60	0.65	0.67	0.57
		Stability	**0.63**	**0.72**	**0.72**	**0.60**
	128	Pred. Prob.	0.75	0.74	0.65	0.75
		Drop-Out	0.85	0.78	0.83	0.75
		Stability	**0.88**	**0.90**	**0.84**	**0.79**
	512	Pred. Prob.	0.78	0.68	0.42	0.45
		Drop-Out	0.78	0.78	0.42	0.45
		Stability	**0.79**	**0.71**	**0.78**	**0.68**
German Credit	64	Pred. Prob.	0.27	0.04	0.27	0.17
		Drop-Out	0.73	0.45	0.60	0.17
		Stability	**0.77**	**0.67**	**0.78**	**0.76**
	128	Pred. Prob.	0.85	0.76	0.85	0.91
		Drop-Out	0.86	0.91	0.85	0.91
		Stability	**0.89**	**0.91**	**0.89**	**0.92**
	512	Pred. Prob.	0.42	0.29	0.27	0.19
		Drop-Out	0.43	0.36	0.28	0.33
		Stability	**0.61**	**0.60**	**0.67**	**0.69**
Diabetes	64	Pred. Prob.	0.09	0.04	0.27	0.23
		Drop-Out	0.24	0.41	**0.54**	**0.50**
		Stability	**0.27**	**0.55**	0.31	0.25
	128	Pred. Prob.	0.16	0.06	0.17	0.16
		Drop-Out	0.46	0.55	**0.54**	**0.63**
		Stability	**0.52**	**0.57**	0.44	0.52
	512	Pred. Prob.	0.61	0.35	0.12	0.19
		Drop-Out	0.71	0.42	**0.42**	**0.51**
		Stability	**0.79**	**0.40**	0.39	0.40
Bank	64	Pred. Prob.	0.26	0.04	0.27	0.17
		Drop-Out	0.24	0.60	0.60	**0.60**
		Stability	**0.77**	**0.67**	**0.78**	0.76
	128	Pred. Prob.	0.45	0.54	0.73	0.62
		Drop-Out	0.62	0.70	0.75	**0.82**
		Stability	**0.89**	**0.71**	**0.78**	0.84
	512	Pred. Prob.	0.42	0.29	0.27	0.11
		Drop-Out	0.44	0.29	**0.37**	**0.43**
		Stability	**0.61**	**0.60**	0.30	0.38

Table B.5: Correlation between the proposed consistency measures and various multiplicity evaluation metrics. New additional measures: $S_1(\mathbf{x}) = \frac{1}{k}\sum|f(\mathbf{x}_i) - f(\mathbf{x})|$ (absolute variability). $S_2(\mathbf{x}) = \frac{1}{k}\sum(f(\mathbf{x}_i) - f(\mathbf{x}))^2$ (squared variability). The proposed Stability measure outperforms both dropout-based method and purely variability-based alternatives

Dataset	Number of Shots	Measure	Arbitrariness	Pairwise Disagreement	Prediction Variance	Prediction Range
Adult	128	Pred. Prob.	0.67	0.62	0.30	0.54
		Drop-Out	0.74	0.83	0.69	0.81
		$S_1(\mathbf{x})$	0.70	0.65	0.63	0.72
		$S_2(\mathbf{x})$	0.70	0.64	0.60	0.73
		Stability	**0.80**	**0.96**	**0.84**	**0.91**

Table B.6: Ablation study on different k values: Correlation between our stability measure (evaluated on a single model) and various measures of multiplicity for different sample sizes k on the Diabetes dataset (T0 model). We observe better performance with increasing k as suggested by our theoretical results. Larger sample size k values are advantageous, as they ensure that the guarantees hold with high probability. However, computational cost of model inference increases.

k	**Prediction Range**	**Prediction Variance**	**Pairwise Disagreement**	**Arbitrariness**
2	0.77	0.77	0.53	0.52
5	0.82	0.83	0.56	0.55
10	0.87	0.87	0.62	0.61
20	**0.89**	**0.88**	**0.70**	**0.79**

Table B.7: Ablation study on different σ values: Correlation between our stability measure (evaluated on one model) and various evaluation measures for different values of σ and evaluated multiplicity for Diabetes dataset and 128-shot case (T0 model). Best performance observed when $\sigma = 10^{-2}$. To guide the choice of σ, one could consider the spread of training data points in the embedding space (e.g., we use a value equivalent to 10% of the variance of the training data). For all our experiments, we used a fixed value of 0.01, which consistently worked well across different datasets and experiments. When σ is too small, we basically sample (almost) the same points and our stability measure is not more informative than the prediction probability. When σ is too large, one loses all information about the data point.

σ	**Prediction Range**	**Prediction Variance**	**Pairwise Disagreement**	**Arbitrariness**
10^{-4}	0.82	0.83	0.84	0.80
10^{-3}	0.91	0.92	0.90	0.86
10^{-2}	**0.95**	**0.93**	**0.95**	**0.92**
10^{-1}	0.10	0.08	0.33	0.23

Table B.8: This table reports the correlation between the stability measure and various evaluated multiplicity for the 512-shot setting on the Diabetes dataset. The stability measure $S_{k,\sigma}(x, f)$ shows a higher correlation with multiplicity compared to predicted probabilities and drop-out and ensemble method, indicating that the stability measure $S_{k,\sigma}(x, f)$ better informs multiplicity than the other measures.

Method	**Arbitrariness**	**Pairwise Disagreement**	**Prediction Variance**	**Prediction Range**
Pred. Prob.	0.21	0.23	0.24	0.30
drop-out $p = 0.01$	0.21	0.23	0.27	0.28
drop-out $p = 0.1$	0.62	0.61	0.59	0.64
drop-out $p = 0.2$	0.74	0.36	0.53	0.54
drop-out $p = 0.5$	0.16	0.17	0.18	0.16
Stability	**0.80**	**0.89**	**0.74**	**0.68**

Table B.9: We compare the correlation in our default LoRA setting against Prompt Tuning and Prefix Tuning (Bank dataset) to assess the generalizability of our method beyond LoRA. Although the stability measure achieves the highest correlations under LoRA, it still provides meaningful signals under Prompt and Prefix tuning.

Dataset	Measure	Arbitrariness	Pairwise Disagreement	Prediction Variance	Prediction Range
Bank	Pred. Prob. (LoRA)	**0.54**	**0.57**	**0.73**	**0.62**
	Pred. Prob. (Prompt Tuning)	0.50	0.48	0.61	0.55
	Pred. Prob. (Prefix Tuning)	0.52	0.49	0.59	0.51
	Drop-Out (LoRA)	**0.62**	**0.70**	**0.75**	**0.51**
	Drop-Out (Prompt Tuning)	0.48	0.53	0.60	0.49
	Drop-Out (Prefix Tuning)	0.55	0.50	0.58	0.46
	Stability (LoRA)	**0.79**	**0.84**	**0.87**	**0.86**
	Stability (Prompt Tuning)	0.63	0.60	0.58	0.61
	Stability (Prefix Tuning)	0.59	0.62	0.60	0.57

Appendix C

C.1 Background on BLEU metric

The BLEU (Bilingual Evaluation Understudy [320] score is a standard metric for assessing the quality of machine translation. It quantifies the degree of overlap between a system-generated translation and one or more human reference translations. The score relies on modified n-gram precision ($n \in \{1, 2, 3, 4\}$), together with a brevity penalty (BP) that discourages excessively short outputs:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \quad \text{BP} = \begin{cases} 1 & \text{if } c > r, \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r, \end{cases}$$

where p_n denotes the modified n-gram precision, w_n are the associated weights, c is the length of the candidate translation, and r is the length of the closest reference. When multiple references are given, BLEU counts n-gram matches against all references and uses the maximum match count for each n-gram. Each n-gram level in BLEU captures progressively deeper aspects of linguistic quality. Unigrams (n=1) assess word choice or adequacy, indicating whether the candidate includes the correct content words. Bigrams (n=2) reflect local fluency by capturing short-range word ordering. Trigrams (n=3) capture phrase-level coherence. 4-grams (n=4) enforce sentence-level fluency by requiring longer, contiguous sequences to match the reference.

C.2 Prompt Templates

Generating paraphrased and semantically equivalent queries

For each query q_0, we use LLaMA-3.1-70B to generate n paraphrases $\mathcal{P}(q_0) = \{p_1, \ldots, p_n\}$. To ensure answerability, we provide the ground truth answer as part of the prompt and instruct the model to generate paraphrases that preserve the exact meaning such that each paraphrase can be answered in the same way. This allows us to simulate semantically equivalent inputs without altering the expected outputs. See prompt used for short-form and long form QA tasks below:

Paraphrasing -- Short-form QA

You are given an input sentence. Your task is to generate `n` diverse paraphrases of this sentence. You can paraphrase by using synonyms, changing sentence structure, or rephrasing in any other way, but each paraphrase should preserve the original meaning. Each paraphrase you create must be answerable by the exact same answer provided below.
Format your output as follows:
`<paraphrase1>` paraphrased sentence 1 `</paraphrase1>`
`<paraphrase2>` paraphrased sentence 2 `</paraphrase2>`
...
`<paraphrasen>` paraphrased sentence `n` `</paraphrasen>`
Input sentence: {`sentence`} **Required answer:** {`answer`}
Please return only the paraphrases in the specified format.

Paraphrasing -- Long-form QA

You are given an input question sentence. Your task is to generate `n` diverse paraphrases of this question. You can paraphrase by using synonyms, changing sentence structure, or rephrasing in any other way, but each paraphrase should preserve the original question meaning and lead to similar answers.
Format your output as follows:
`<paraphrase1>` paraphrased sentence 1 `</paraphrase1>`
`<paraphrase2>` paraphrased sentence 2 `</paraphrase2>`
...
`<paraphrasen>` paraphrased sentence `n` `</paraphrasen>`
Input question: {`sentence`}
Please return only the paraphrases in the specified format.

LLM-Judge Consistency Evaluation

To assess semantic consistency between generated outputs across paraphrased queries, we employ LLM-based pairwise judgments as part of our evaluation pipeline. These judgments are designed to determine whether two answers convey the same core information, even if they differ in surface form. For all `LLaMA-3.1-8B` experiments, we use `LLaMA 3.3 70B` as the evaluator. For all `Qwen-2.5-3B` experiments, we use `GPT-4o`, an external closed-source model, as the evaluator. An example evaluation prompt is shown below:

Consistency Evaluation -- Long-form QA

You are an evaluator assessing two different answers that were generated for the same question. Your task is to determine if the two answers are consistent with each other.
Consider them consistent if they present the same core information about the underlying question. Consider them inconsistent if they provide different information, present different facts, or address the underlying question in different ways.
Based on this, reply with only `yes` or `no`. Do not output anything else.
Answer 1: {`output_i`}
Answer 2: {`output_j`}
Are these two answers consistent? (yes/no). Response:

Consistency Evaluation -- Short-form QA

Given the following two outputs sentences, reply with `yes` if the two sentences are consistent with each other, or `no` if they are not. Do not output anything else.
Sentence 1: {`output_i`}
Sentence 2: {`output_j`}
Are these sentences consistent? (yes/no). Response:

C.3 Expanded Experiments

Table C.1: **Accuracy across datasets and query variants (LLaMA-3.1-8B).** We report accuracy for original queries, synthetically generated paraphrased queries, and paraphrased queries with fixed retrieval. Across all settings, accuracy remains relatively similar, indicating that paraphrasing and retrieval shifts have limited effect on final answer correctness on average. See result for `Qwen-2.5-3B` model in Table C.3.

Short-form & Multi-hop QA: Accuracy (%)

Dataset	Original Queries			Paraphrased Queries			Paraphrased (Fixed Docs)		
	EM	F1	RM	EM	F1	RM	EM	F1	RM
TriviaQA	56.0	66.1	74.0	55.0	64.4	73.3	58.7	67.3	75.0
HotpotQA	37.0	44.1	42.0	36.4	43.5	42.4	33.7	40.7	39.4
2Wiki	28.0	33.9	37.0	25.9	31.3	32.7	26.9	31.7	33.3
MuSiQue	8.0	15.3	12.0	8.3	14.1	11.0	11.0	17.5	15.0

Long-form QA: Accuracy (%)

Dataset	Original Queries		Paraphrased Queries		Paraphrased (Fixed Docs)	
	ROUGE	LLM-Acc	ROUGE	LLM-Acc	ROUGE	LLM-Acc
`ELI5`	21.9	74.0	20.7	71.3	20.8	70.3

Table C.2: **Disentangling sources of inconsistency in RAG systems (`Qwen-2.5-3B`).** Retriever consistency is low across datasets, suggesting that paraphrased queries often retrieve non-overlapping documents. This introduces context variability that is reflected in the end-to-end consistency scores. Fixing retrieval improves consistency, but variation remains, revealing the generator's sensitivity to input phrasing even with identical evidence.

Dataset	End-to-End Consistency		Generator (LLM) Consistency		Retriever Consistency
	Lexical	LLM-Judge	Lexical	LLM-Judge	Jaccard Overlap
TriviaQA	47.9	73.0	58.6	87.5	32.5
HotpotQA	32.7	63.6	48.0	77.3	46.0
2Wiki	32.3	62.6	44.6	70.7	52.4
MuSiQue	25.7	49.5	45.7	67.3	36.6
`Eli5`	6.6	35.3	14.4	62.3	27.1

Table C.3: **Accuracy across datasets and query variants (`Qwen-2.5-3B`).** We report accuracy for original queries, synthetically generated paraphrased queries, and paraphrased queries with fixed retrieval. Across all settings, accuracy remains relatively similar, indicating that paraphrasing and retrieval shifts have limited effect on final answer correctness on average.

Short-form & Multi-hop QA: Accuracy (%)									
Dataset	**Original Queries**			**Paraphrased Queries**			**Paraphrased (Fixed Docs)**		
	EM	F1	RM	EM	F1	RM	EM	F1	RM
TriviaQA	42.0	50.7	58.0	46.3	54.1	64.3	43.0	51.1	62.3
HotpotQA	20.0	28.3	37.0	20.9	27.7	38.4	18.2	26.4	38.4
2Wiki	13.0	20.4	36.0	11.1	19.8	34.7	12.5	20.2	32.3
MuSiQue	4.0	10.0	9.0	6.0	9.9	8.0	5.3	9.7	7.7

Long-form QA: Accuracy (%)						
Dataset	**Original Queries**		**Paraphrased Queries**		**Paraphrased (Fixed Docs)**	
	ROUGE	LLM-Acc	ROUGE	LLM-Acc	ROUGE	LLM-Acc
`ELI5`	22.1	38.0	21.5	37.3	20.8	35.7

Table C.4: **Comparison between Con-RAG vs Baselines (Short-form QA Tasks) (`Qwen-2.5-3B`).** Lexical consistency measured via BLEU score while and information consistency measured using an LLM-judge. Con-RAG is trained with a group-similarity reward plus an accuracy reward (no KL), and consistently yields higher end-to-end and generator-only consistency while also improving accuracy over original queries.

		Accuracy (%)			**End-to-End Consistency (%)**		**Generator (LLM) Consistency (%)**	
Dataset	**Method**	EM	F1	RM	Lexical	Inform.	Lexical	Inform.
TriviaQA	RAG	42.0	50.7	58.0	47.9	73.0	58.6	87.5
	DRAG	42.0	50.7	58.0	47.9	73.5	58.6	84.7
	CoT-RAG	37.0	44.5	61.0	41.1	72.3	52.2	82.3
	SFT	35.0	40.4	43.0	53.3	72.2	73.4	85.0
	Con-RAG	**60.0**	**66.0**	**68.0**	**67.1**	**81.8**	**80.5**	**89.5**
HotpotQA	RAG	20.0	28.3	37.0	32.7	63.6	48.0	77.3
	DRAG	20.0	28.3	37.0	32.7	64.3	48.0	76.8
	CoT-RAG	29.0	32.8	37.0	28.5	63.6	35.7	71.2
	SFT	30.0	35.4	32.0	63.3	77.1	74.5	85.7
	Con-RAG	**36.0**	**43.1**	**38.0**	**64.6**	**78.2**	**77.8**	**86.7**
MuSiQue	RAG	4.0	10.0	9.0	25.7	49.5	45.7	67.3
	DRAG	4.0	10.0	9.0	25.7	50.3	45.7	69.2
	CoT-RAG	5.0	10.9	9.0	18.1	52.0	26.5	62.0
	SFT	25.0	30.6	27.0	57.7	65.3	69.8	77.2
	Con-RAG	**27.0**	**31.9**	**28.1**	**69.8**	**70.1**	**70.4**	**82.0**
2Wiki	RAG	13.0	20.4	36.0	32.3	62.6	44.6	70.7
	DRAG	13.0	20.4	36.0	32.3	63.0	44.6	70.9
	CoT-RAG	23.0	27.0	30.0	23.5	62.3	32.7	67.0
	SFT	**37.0**	**38.9**	**38.0**	**70.9**	75.8	**84.8**	86.9
	Con-RAG	**37.0**	38.4	37.0	68.2	**76.6**	**84.8**	**89.1**

Table C.5: **Comparison between Con-RAG vs. Baselines (Long-form QA Task).** Con-RAG is trained using only the group-similarity reward with a small KL regularizer (no accuracy supervision). Despite no ground-truth, it achieves the best end-to-end and generator consistency and also improves answer quality over baselines, whereas SFT on reference answers underperforms in this open-ended setting (`Qwen-2.5-3B`).

		Accuracy (%)		**End-to-End Consistency (%)**		**Generator (LLM) Consistency (%)**	
Dataset	**Method**	ROUGE	LLM-Acc	Lexical	Inform.	Lexical	Inform.
	RAG	22.1	38.0	6.6	35.3	14.4	62.3
	DRAG	22.1	38.0	6.6	35.3	14.4	63.8
ELI5	CoT-RAG	21.1	36.0	4.9	34.0	9.6	55.5
	SFT	**24.3**	36.0	5.4	17.2	7.0	19.0
	Con-RAG	22.6	**58.0**	**9.3**	**42.8**	**17.9**	**67.5**

Table C.6: **Effect of Inference Temperature on Standard RAG(`ELI5 - Qwen-2.5-3B`).** We vary only the decoding temperature T at inference to study its effect on consistency and accuracy. Moderate temperature ($T = 0.5$) improves LLM agreement and lexical consistency compared to deterministic decoding ($T = 0.0$), while preserving accuracy. However, higher temperatures ($T \geq 1.0$) degrade both consistency and accuracy, with outputs at $T = 2.0$ nearly collapsing.

T	**Accuracy (%)**		**End-to-End Cons. (%)**		**Generator Cons. (%)**	
	ROUGE	LLM-Acc	Lexical	LLM-Judge	Lexical	LLM-Judge
0.0	22.1	38.0	6.6	35.3	14.4	62.3
0.5	**21.4**	**52.0**	**10.4**	**37.7**	**15.2**	**65.3**
1.0	21.8	48.0	2.5	34.0	5.2	59.5
2.0	6.1	0.0	0.1	2.0	0.2	1.5

Appendix D

D.1 Background on Fisher Information Matrix

Definition D.1 (Positive Semi-definite Matrices)**.** *A matrix $A \in \mathbb{R}^{d\times d}$ is said to be positive semi-definite if it is symmetric and for all non-zero vectors $\mathbf{x} \in \mathbb{R}^d$, the following condition holds:*

$$\mathbf{x}^T A\mathbf{x} \geq 0 \quad \textit{for all} \quad \mathbf{x} \in \mathbb{R}^d.$$

The eigenvalues of a positive semi-definite are non-negative, i.e., $\lambda_i(A) \geq 0$ for all eigenvalues λ_i.

Definition D.2 (Löwner Order)**.** *Let $A, B \in \mathbb{R}^{d\times d}$ be symmetric matrices. We say that A is greater than or equal to B in the Löwner order, denoted $A \succeq B$, if and only if the matrix $A - B$ is positive semi-definite. That is,*

$$A \succeq B \quad \textit{if and only if} \quad x^T(A-B)x \geq 0 \quad \textit{for all} \quad x \in \mathbb{R}^d.$$

If $A \succ B$, then $A - B$ is positive definite, meaning A is strictly greater than B in the Löwner order.

Lemma D.1 (Trace Inequality for Positive Semi-definite Matrices)**.** *For positive semi-definite matrices $A, B \in \mathbb{R}^{d\times d}$ where $A \succ B$, then:*

$$\textit{Tr}(A^{-1}) < \textit{Tr}(B^{-1})$$

Proof. Since $A \succ B$, we have $B^{-1} \succ A^{-1}$ by the Löwner order inversion property. The trace operator preserves this inequality because for any $X \succ Y \succ 0$:

$$\text{Tr}(X) = \sum_{i=1}^{d} \lambda_i(X) > \sum_{i=1}^{d} \lambda_i(Y) = \text{Tr}(Y)$$

where $\lambda_i(\cdot)$ denotes eigenvalues in descending order. □

Definition 5.2 (Fisher Information Matrix [217])**.** *Let $\mathcal{L}(\theta)$ be the log-likelihood of a parametric distribution $p(y, x; \theta)$, where θ is the parameter vector to be estimated. The* Fisher Information Matrix *(FIM) at parameter θ is defined as:*

$$\mathcal{I}(\theta) = \mathbb{E}_{\mathbf{x},y}\left[\nabla_\theta \log p(y, \mathbf{x}; \theta)\, \nabla_\theta \log p(y, \mathbf{x}; \theta)^\top\right].$$

Fisher information captures the amount of information that an observable random variable x carries about an unknown parameter θ of a distribution that models x. We use the notation $\mathcal{I}(\theta; y, \mathbf{x})$ to denote the Fisher information about θ carried by single observation $y, \mathbf{x}$.

D.2 Proof of Theorem 5.1

Theorem 5.1 (CFEs Improve Model Parameter Estimation)**.** *Let $\mathbf{w}_s$ and $\mathbf{w}_s^{(\text{cf})}$ be the student parameters obtained via MLE on $\mathcal{D}$ (standard) and $\mathcal{D}_{\text{cf}}$ (CFE-infused). Assuming the teacher's parameters $\mathbf{w}_t$ capture the true data-generating distribution, that CFEs lie near the decision boundary, and that the second moments $\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top] \approx \mathbb{E}_{\mathbf{x}_c}[\mathbf{x}_c\mathbf{x}_c^\top]$. Then estimation error satisfies:*

$$\mathbb{E}\left[\|\mathbf{w}_s^{(\text{cf})} - \mathbf{w}_t\|^2\right] < \mathbb{E}\left[\|\mathbf{w}_s - \mathbf{w}_t\|^2\right].$$

Proof. For a single observation $(\mathbf{x}, y)$, the log-likelihood is:

$$\log p(y|\mathbf{x}; \mathbf{w}) = y \log \sigma(\mathbf{w}^\top \mathbf{x}) + (1 - y) \log(1 - \sigma(\mathbf{w}^\top \mathbf{x})) \tag{D.1}$$

Taking the gradient with respect to $\mathbf{w}$:

$$\nabla_{\mathbf{w}} \log p(y|\mathbf{x}; \mathbf{w}) = (y - \sigma(\mathbf{w}^\top \mathbf{x}))\mathbf{x} \tag{D.2}$$

To prove Theorem 5.1, we first (1) Characterize the Fisher information for individual observations, (2) Establish asymptotic normality of MLE, (3) Compare information matrices of standard vs. CFE-infused datasets, and (4) Apply trace inequality to connect information to estimation error.

(1) Fisher Information for Logistic Regression: For a logistic regression model with parameters $\mathbf{w}$, Lets denote Fisher Information Matrix (FIM) for observations $y, \mathbf{x}$ as:

$$\mathcal{I}(\mathbf{w}; y, \mathbf{x}) = \mathbb{E}_{y,\mathbf{x}} \left[\nabla_{\mathbf{w}} \log p(y, \mathbf{x}; \mathbf{w}) \nabla_{\mathbf{w}} \log p(y, \mathbf{x}; \mathbf{w})^\top \right] \tag{D.3}$$

$$\nabla_{\mathbf{w}} \log p(y, \mathbf{x}; \mathbf{w}) = \nabla_{\mathbf{w}} \log p(y|\mathbf{x}; \mathbf{w}) + \underbrace{\nabla_{\mathbf{w}} \log p(\mathbf{x})}_{=0}. \tag{D.4}$$

The gradient of $\log p(\mathbf{x})$ is zero because $p(\mathbf{x})$ is independent of the model parameters $\mathbf{w}$.

Using the law of total expectation:

$$\mathcal{I}(\mathbf{w}; y, \mathbf{x}) = \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}} \left[\nabla_{\mathbf{w}} \log p(y|\mathbf{x}; \mathbf{w}) \nabla_{\mathbf{w}} \log p(y|\mathbf{x}; \mathbf{w})^\top \right] \tag{D.5}$$

Substituting Equation D.2:

$$\mathcal{I}(\mathbf{w}; y, \mathbf{x}) = \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{y|\mathbf{x}} [(y - \sigma(\mathbf{w}^\top \mathbf{x}))^2 \mathbf{x}\mathbf{x}^\top] \right] \tag{D.6}$$

$$= \mathbb{E}_{\mathbf{x}} \left[\mathbf{x}\mathbf{x}^\top \mathbb{E}_{y|\mathbf{x}} [(y - \sigma(\mathbf{w}^\top \mathbf{x}))^2] \right] \tag{D.7}$$

The term $\mathbb{E}_{y|\mathbf{x}} \left[\left(y - \sigma(\mathbf{w}^\top \mathbf{x}) \right)^2 \right]$ is the variance of $y|\mathbf{x}$. Where $y|\mathbf{x} \sim \text{Bernoulli}(\sigma(\mathbf{w}^\top \mathbf{x}))$, we compute:

$$\mathbb{E}_{y|\mathbf{x}} [(y - \sigma(\mathbf{w}^\top \mathbf{x}))^2] = \text{Var}(y|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x})(1 - \sigma(\mathbf{w}^\top \mathbf{x})) \tag{D.8}$$

Thus:

$$\mathcal{I}(\mathbf{w}; y, \mathbf{x}) = \mathbb{E}_{\mathbf{x}} [\sigma(\mathbf{w}^\top \mathbf{x})(1 - \sigma(\mathbf{w}^\top \mathbf{x})) \mathbf{x}\mathbf{x}^\top] \tag{D.9}$$

The variance term is maximized when $\mathbf{w}^\top \mathbf{x} = 0$ (i.e., at the decision boundary), where it equals 0.25.

(2) Asymptotic Distribution of MLE: Under regularity conditions [233], the MLE estimator satisfies:

$$\sqrt{k}(\mathbf{w}_s - \mathbf{w}_t) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\mathbf{w}_t; \mathcal{D})) \tag{D.10}$$

where $\mathcal{I}(\mathbf{w}_t; \mathcal{D}) = \sum_{i=1}^{k} \mathcal{I}(\mathbf{w}_t; y_i, \mathbf{x}_i)$ is the total Fisher information of k independent observations of $y_i, \mathbf{x}_i$ (Additivity property of fisher information [321]).

The mean squared error (MSE) [322] decomposes as:

$$\mathbb{E}\|\mathbf{w}_s - \mathbf{w}_t\|^2 = \underbrace{\text{Tr}(\text{Cov}(\mathbf{w}_s))}_{\text{Variance}} + \underbrace{\|\text{Bias}(\mathbf{w}_s)\|^2}_{\text{Bias}} \tag{D.11}$$

For MLE, $\text{Bias}(\mathbf{w}_s) \to 0$ as $k \to \infty$, so: $\mathbb{E}\|\mathbf{w}_s - \mathbf{w}_t\|^2 \approx \text{Tr}(\mathcal{I}^{-1}(\mathbf{w}_t; \mathcal{D}))$

The next step of the proof we compare the fisher information between a standard dataset and

CFE-infused dataset. Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{k}$ be a dataset of k standard samples, and let $\mathcal{D}_{\text{cf}} = \{\mathbf{x}_i\}_{i=1}^{k/2} \cup \{\mathbf{x}_{c_j}\}_{j=1}^{k/2}$ be an CFE-infused dataset containing $k/2$ standard samples and $k/2$ CFEs.

Standard Samples: Far from decision boundary $\Rightarrow \mathbf{w}_t^\top \mathbf{x}_i \gg 0$ or $\ll 0$. Thus:

$$\sigma(\mathbf{w}_t^\top \mathbf{x}_i)(1 - \sigma(\mathbf{w}_t^\top \mathbf{x}_i)) = \epsilon_i < 0.25 \tag{D.12}$$

Their FIM contribution is: $\mathcal{I}(\mathbf{w}_t; \mathbf{x}_i) = \mathbb{E}_{\mathbf{x}}[\epsilon_i \mathbf{x}_i \mathbf{x}_i^\top]$.

CFE Samples: Near boundary $\Rightarrow \mathbf{w}_t^\top \mathbf{x}_c = 0 \Rightarrow \sigma(0) = 0.5$. Thus:

$$\sigma(\mathbf{w}_t^\top \mathbf{x}_c)(1 - \sigma(\mathbf{w}_t^\top \mathbf{x}_c)) = 0.25 \tag{D.13}$$

Their FIM contribution is maximal: $\mathcal{I}(\mathbf{w}_t; \mathbf{x}_c) = \mathbb{E}_{\mathbf{x}}[0.25 \mathbf{x}_c \mathbf{x}_c^\top]$.

Since $\mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^\top] \approx \mathbb{E}_{\mathbf{x}_c}[\mathbf{x}_c \mathbf{x}_c^\top]$ and $0.25 \gg \epsilon_i$, we have $\mathcal{I}(\mathbf{w}_t; \mathcal{D}_{cf}) \succ \mathcal{I}(\mathbf{w}_t; \mathcal{D})$ in the Löwner order (see Definition D.2).

Remark D.1 (Feature Spanning)**.** *Note that for logistic regression the feature vector is augmented with the parameter bias term, i.e.,* $\mathbf{x} = [1, \tilde{\mathbf{x}}^\top]^\top$*, hence,the outer product* $\mathbf{x}\mathbf{x}^\top$ *has a non-zero norm. The first element of* $\mathbf{x}$ *is always* 1*, ensuring* $\|\mathbf{x}\|^2 \geq 1$*. Thus,* $\mathbf{x}\mathbf{x}^\top$ *cannot be the zero matrix, even if* $\tilde{\mathbf{x}} = \mathbf{0}$*. This guarantees that each CFE example* $\mathbf{x}_c$ *contributes a non-degenerate rank-1 term to the FIM.*

The final step leverages the trace inequality for covariance matrices (see Lemma D.1). If $\mathcal{I}(\mathbf{w}_t; \mathcal{D}_{cf}) \succ \mathcal{I}(\mathbf{w}_t; \mathcal{D})$ then $\text{Tr}(\mathcal{I}^{-1}(\mathbf{w}_t; \mathcal{D}_{cf})) < \text{Tr}(\mathcal{I}^{-1}(\mathbf{w}_t; \mathcal{D}))$. Thus, CFE infusion reduces estimation error:

$$\mathbb{E}\left[\|\mathbf{w}_s^{(\text{cf})} - \mathbf{w}_t\|^2\right] < \mathbb{E}\left[\|\mathbf{w}_s - \mathbf{w}_t\|^2\right] \tag{D.14}$$

Remark D.2 (Datapoint Diversity)**.** *For the total FIM $\mathcal{I}(\mathbf{w}_t; \mathcal{D}_{cf})$ to be invertible, the set of feature vectors $\{\mathbf{x}_i\}$ must span $\mathbb{R}^d$ which will hold if we have enough samples.*

□

D.3 Background on Hausdorff Distance

This section provides definitions and geometric preliminaries needed for proofs of Lemma 5.1 and Theorem 5.2.

Definition D.3 (Line Segment)**.** *Let $\mathbf{x}_i, \mathbf{x}_i' \in \mathbb{R}^{n \times d}$ be two points in the $n \times d$ space. The line segment $[\mathbf{x}_i, \mathbf{x}_i']$ connecting $\mathbf{x}_i$ and $\mathbf{x}_i'$ is defined as the set of points $\gamma(\lambda)$ for $\lambda \in [0, 1]$, where*

$$\gamma(\lambda) = (1 - \lambda)\mathbf{x}_i + \lambda \mathbf{x}_i', \quad \lambda \in [0, 1].$$

This defines all the points on a space between $\mathbf{x}_i$ and $\mathbf{x}_i'$ in $\mathbb{R}^{n \times d}$.

Lemma D.2 (Intermediate Value Theorem)**.** *Let $f : [a, b] \to \mathbb{R}$ be a continuous function, and let $f(a) \neq f(b)$. If y is any value between $f(a)$ and $f(b)$, then there exists $c \in (a, b)$ such that $f(c) = y$.*

Definition 5.3 (Hausdorff Distance)**.** *Let $\mathcal{M}_t, \mathcal{M}_s \subseteq \mathbb{R}^{n \times d}$ be two non-empty subsets of a metric space. The* Hausdorff distance *is defined as:*

$$\mathrm{H}(\mathcal{M}_s, \mathcal{M}_t) = \max\Big\{ \sup_{\mathbf{x} \in \mathcal{M}_t} \inf_{\mathbf{u} \in \mathcal{M}_s} \|\mathbf{x} - \mathbf{u}\|_F, \ \sup_{\mathbf{u} \in \mathcal{M}_s} \inf_{\mathbf{x} \in \mathcal{M}_t} \|\mathbf{u} - \mathbf{x}\|_F \Big\}.$$

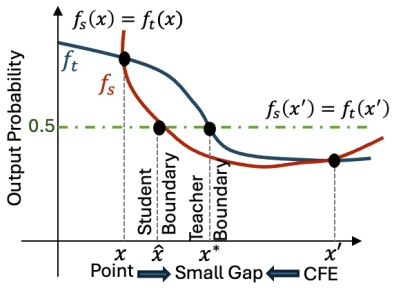


Figure D.1: Intuition for Theorem 5.2

D.4 Proof of Theorem 5.2

Lemma 5.1 (Existence of Boundary Crossing for Counterfactual Pairs)**.** *Let* $f_t : \mathbb{R}^{n\times d} \to [0,1]$ *be a continuous function. For a datapoint and its counterfactual pair* $(\mathbf{x}_i, \mathbf{x}_i')$*, there exists a point* $\mathbf{x}_i^\star = \alpha\mathbf{x}_i + (1-\alpha)\mathbf{x}_i'$ *for an* $\alpha \in (0,1)$ *(on the line joining* $\mathbf{x}_i$ *and* $\mathbf{x}_i'$*) such that:* $f_t(\mathbf{x}_i^\star) = 0.5$*.*

Proof. Define a line segment from x_i to x_i' using a parameterization: $\gamma(\lambda) = (1-\lambda)x_i + \lambda x_i'$ for $\lambda \in [0,1]$. This defines a continuous path from x_i to x_i' in $\mathbb{R}^d$. Now define the real-valued function $g : [0,1] \to \mathbb{R}$ by: $g(\lambda) = f_t(\gamma(\lambda)) = f_t((1-\lambda)x_i + \lambda x_i')$.

Since f_t is continuous on $\mathbb{R}^d$, and $\gamma(\lambda)$ is continuous in λ, the composition $g(\lambda)$ is continuous on the closed interval $[0,1]$.

Now, evaluate the endpoints of this function: $g(0) = f_t(x_i) < 0.5, g(1) = f_t(x_i') > 0.5$.

Thus, we have $g(0) < 0.5 < g(1)$, and by the Intermediate Value Theorem (see Lemma D.2), since g is continuous on $[0,1]$, there exists $\lambda^\star \in (0,1)$ such that: $g(\lambda^\star) = 0.5$.

Define $x_i^\star = \gamma(\lambda^\star) = (1-\lambda^\star)x_i + \lambda^\star x_i' \in [x_i, x_i']$. Then: $f_t(x_i^\star) = g(\lambda^\star) = 0.5$.

Hence, the point $x_i^\star \in [x_i, x_i']$ lies on the segment and satisfies $f_t(x_i^\star) = 0.5$, as required. □

Theorem 5.2 (Teacher–Student Boundary Proximity)**.** *Let* $f_t,\ f_s : \mathbb{R}^{n\times d} \to; [0,1]$ *be the teacher and student model, with decision boundaries* $\mathcal{M}_t = \{\mathbf{x} \mid f_t(\mathbf{x}) = 0.5\}$ *and* $\mathcal{M}_s = \{\mathbf{x} \mid f_s(\mathbf{x}) = 0.5\}$*,*

respectively. Assume we observe a CFE-infused dataset $\mathcal{D}_{cf} = \{(\mathbf{x}_i, \mathbf{x}_i')\}_{i=1}^{k}$ *satisfying, for every pair* $(\mathbf{x}_i, \mathbf{x}_i')$*: (A1) Minimal perturbation:* $\|\mathbf{x}_i - \mathbf{x}_i'\|_F \leq \alpha$ *with* $\alpha > 0$*; (A2) Exact distillation:* $f_s(\mathbf{x}_i) = f_t(\mathbf{x}_i)$ *and* $f_s(\mathbf{x}_i') = f_t(\mathbf{x}_i')$*; and (A3)* ε*-spread along the teacher and student boundary, i.e., for each pair, there exist a teacher's crossing point* $\mathbf{x}_i^\star = \alpha\mathbf{x}_i + (1-\alpha)\mathbf{x}_i'$ *for* $\alpha \in (0,1)$ *such that* $f_t(x_i^\star) = 0.5$ *and for every* $a \in \mathcal{M}_t$*, there exists an* i *with* $\|a - \mathbf{x}_i^\star\|_2 \leq \varepsilon$*. Then the Hausdorff distance between the decision boundaries obeys:* $\mathrm{H}(\mathcal{M}_s, \mathcal{M}_t) \leq \alpha + \varepsilon$.

Proof. To prove Theorem 5.2, we bound the Hausdorff distance between the student's and teacher's decision boundaries using the given assumptions. We bound each term separately of the Hausdorff distance (see Defintion 5.3).

We first bound $\sup_{\mathbf{x}\in\mathcal{M}_t} \inf_{\mathbf{u}\in\mathcal{M}_s} \|\mathbf{x} - \mathbf{u}\|_F$:

For any $a \in \mathcal{M}_t$, by assumption (A3), there exists a CFE pair $(\mathbf{x}_i, \mathbf{x}_i')$ with teacher crossing point $\mathbf{x}_i^\star \in \mathcal{M}_t$ such that:

$$\|a - \mathbf{x}_i^\star\|_F \leq \varepsilon. \tag{D.15}$$

The segment $[\mathbf{x}_i, \mathbf{x}_i']$ has length $\|\mathbf{x}_i - \mathbf{x}_i'\|_F \leq \alpha$ (A1). By Lemma 5.1 and (A2) Exact distillation, the student's boundary $\mathcal{M}_s$ intersects $[\mathbf{x}_i, \mathbf{x}_i']$ at some $\mathbf{u}_i^\star \in \mathcal{M}_s$. Since $\mathbf{x}_i^\star$ and $\mathbf{u}_i^\star$ lie on $[\mathbf{x}_i, \mathbf{x}_i']$, their distance satisfies:

$$\|\mathbf{x}_i^\star - \mathbf{u}_i^\star\|_F \leq \|\mathbf{x}_i - \mathbf{x}_i'\|_F \leq \alpha. \tag{D.16}$$

Combining Equation D.16 and D.15:

$$\|a - \mathbf{u}_i^\star\|_F \leq \|a - \mathbf{x}_i^\star\|_F + \|\mathbf{x}_i^\star - \mathbf{u}_i^\star\|_F \leq \varepsilon + \alpha. \tag{D.17}$$

Thus, $\inf_{\mathbf{u}\in\mathcal{M}_s} \|a-\mathbf{u}\|_F \le \varepsilon+\alpha$. Taking the supremum over $a\in\mathcal{M}_t$:

$$\sup_{\mathbf{x}\in\mathcal{M}_t}\inf_{\mathbf{u}\in\mathcal{M}_s}\|\mathbf{x}-\mathbf{u}\|_F \le \varepsilon+\alpha. \tag{D.18}$$

Next we bound $\sup_{\mathbf{u}\in\mathcal{M}_s}\inf_{\mathbf{x}\in\mathcal{M}_t}\|\mathbf{u}-\mathbf{x}\|_F$:

From (A1), the distance between the student's cutpoint $\mathbf{u}_i^*$ and the teacher's cutpoint $\mathbf{x}_i^\star$ satisfies:

$$\|\mathbf{u}_i^*-\mathbf{x}_i^\star\|_F \le \|\mathbf{x}_i-\mathbf{x}_i'\|_F \le \alpha \tag{D.19}$$

For any other $\mathbf{u}\in\mathcal{M}_s$:

$$\|\mathbf{u}-\mathbf{x}_i^\star\|_F \le \|\mathbf{u}-\mathbf{u}_i^*\|_F + \|\mathbf{u}_i^*-\mathbf{x}_i^*\|_F \le \varepsilon+\alpha, \tag{D.20}$$

Assuming the CFE pairs $(\mathbf{x}_i,\mathbf{x}_i')$ intersection points are $\varepsilon-$spread (well spread) along the student decision boundary.

Since $\mathbf{x}_i^\star\in\mathcal{M}_t$, we have:

$$\inf_{\mathbf{x}\in\mathcal{M}_t}\|\mathbf{u}-\mathbf{x}\|_F \le \|\mathbf{u}-\mathbf{x}_i^\star\|_F \le \varepsilon+\alpha. \tag{D.21}$$

Taking the supremum over $\mathbf{u}\in\mathcal{M}_s$:

$$\sup_{\mathbf{u}\in\mathcal{M}_s}\inf_{\mathbf{x}\in\mathcal{M}_t}\|\mathbf{u}-\mathbf{x}\|_F \le \varepsilon+\alpha. \tag{D.22}$$

Combining both bounds, the Hausdorff distance is the maximum of the two suprema:

$$\mathrm{H}(\mathcal{M}_s, \mathcal{M}_t) \leq \max\{\varepsilon + \alpha,\ \varepsilon + \alpha\} = \varepsilon + \alpha \tag{D.23}$$

□

D.5 Expanded Experiments

Datasets Details

We evaluate CoD across six text classification benchmarks that span a range of domains. For each k-shot setup, we sample a balanced subset from the training data, selecting $k/2$ examples per class. All experiments are repeated across 5 random seeds, each with a different sampled subset.

- Yelp [227]: We use the Yelp Review Full dataset, filtering for reviews with at most 250 tokens and discarding neutral labels. Labels are binarized: 1–2 as negative and 4–5 as positive. The processed dataset contains 106,624 training examples, 1,000 for validation, and 7,074 for testing, with a slightly imbalanced class distribution (64% negative).

- IMDB [224]: We retain only reviews with shorter lengths. The original test and unsupervised splits are repurposed as validation and test sets, respectively. The resulting data includes 782 training, 858 validation, and 1,578 test samples, with the test set unlabeled.

- SST2 2 [222]: We use the full GLUE-provided training, validation, and test splits without modification. The train/val sets contain 67,349 and 872 examples, respectively. The test set has 1,821 unlabeled examples.

- CoLA [225]: We adopt the standard GLUE splits of the CoLA dataset, yielding 8,551 training, 1,043 validation, and 1,063 unlabeled test samples. The task is binary classification of linguistic acceptability.

- Sentiment140 [223]: We filter the dataset to exclude neutral tweets. The final dataset includes 1,598,400 training, 1,600 validation, and 359 test examples, with balanced label distributions.

- Amazon Polarity [226]: We select examples with shortest length. The processed data includes 1,111 training and 113 validation samples, with roughly balanced sentiment labels.

Counterfactual Explanation Generation Prompt Templates

We provide prompt templates used for counterfactual explanation generation across datasets. Each prompt instructs the model to minimally modify a given input to flip the class label (e.g., sentiment or grammaticality) while preserving meaning and structure. We used gpt-4o-2024-11-20 [221] for our CFE generation.

SST-2 / IMDB / Sentiment140 / Amazon

You are an AI assistant tasked with generating counterfactual explanations for sentiment analysis.
Given a sentence and its true sentiment label, your goal is to make the minimal necessary change to flip the sentiment while preserving the structure and meaning as much as possible.
For example, if the input is:
Sentence: "I love this movie."
True sentiment: Positive
A suitable counterfactual explanation would be: "I dislike this movie."
Now, generate a counterfactual explanation for the following sentence:
Sentence: {sentence}
True sentiment: {sentiment}
Return only the counterfactual sentence, without any additional information.

Yelp

You are an AI assistant tasked with generating counterfactual explanations for sentiment analysis of Yelp reviews.
Given a sentence (a Yelp review) and its true sentiment label (positive or negative), your goal is to make the minimal necessary change to flip the sentiment while preserving the structure and meaning as much as possible.
For example, if the input is:
Sentence: "This restaurant is fantastic, the food was amazing!"
True sentiment: Positive
A suitable counterfactual explanation would be: "This restaurant is terrible, the food was awful!"
Now, generate a counterfactual explanation for the following sentence:
Sentence: {`sentence`}
True sentiment: {`sentiment`}
Return only the counterfactual sentence, without any additional information.

CoLA

You are an AI assistant tasked with generating counterfactual explanations for grammaticality judgment.
Given a sentence and its true grammaticality label (Acceptable or Unacceptable), your goal is to make the minimal necessary change to flip the grammaticality while preserving the structure and meaning as much as possible.
For example, if the input is:
Sentence: "She is going to the store."
True grammaticality: Acceptable
A suitable counterfactual explanation would be: "She is go to the store."
Now, generate a counterfactual explanation for the following sentence:
Sentence: {`sentence`}
True grammaticality: {`sentiment`}
Return only the counterfactual sentence, without any additional information.

Baselines Details

We compare CoD against three task-aware knowledge distillation methods widely used for distillation. CoD uses $k/2$ original samples and their $k/2$ corresponding CFE (a total of k shots) while the baseline methods are trained on k original samples.

- **Knowledge Distillation (KD)** [228]: A classical distillation approach where the student model learns to mimic the teacher's soft target probabilities using Kullback-Leibler (KL) divergence. This method transfers predictive behavior but does not supervise intermediate representations.

- **Layer-wise Distillation (LWD)** [178]: An extension of KD that additionally aligns the student's intermediate hidden representations with those of the teacher. This is typically done via a mean squared error loss over corresponding layers, encouraging the student to internalize not only the final outputs but also the hierarchical feature representations of the teacher.

- **Task-aware Layer-wise Distillation (TED)** [177]: TED augments LWD with learned neural filters at each layer of both teacher and student models. These filters are trained to select task-relevant information from intermediate representations before computing the distillation loss. This selective transfer enables more effective compression by focusing on information critical to task performance.

Models and Hyperparameters

- **DeBERTa-V3 [190].** We fine-tune the teacher model using DeBERTaV3-base, initialized with a classification head for each target task. For the teacher, we use a dropout rate of 0.1, linear learning rate decay, and train for 8 epochs with a fixed learning rate of 2×10^{-5} and batch sizes of $\{32, 64\}$. Optimization is performed using Adam with $\epsilon = 1 \times 10^{-6}$, $\beta_1 = 0.9$, and $\beta_2 = 0.98$, without weight decay. Mixed-precision training with FP16 is used throughout.

  For distillation, the student is initialized from a pre-trained DeBERTa-v3-small or DeBERTa-v3-xsmall model. We search learning rates in the range $[1 \times 10^{-5}, 5 \times 10^{-5}]$, and use a fixed batch size of 8 in our few-shot experiments. All student models are trained for 10 epochs using Adam with the same optimizer settings as the teacher. For KD and LWD baselines, we set the distillation loss weight to 20. For the TED baseline, we use the same hyperparameters for both the filter training and distillation phases, consistent with [177].

- **Qwen2.5 [167].** We use `Qwen/Qwen2.5-1.5B` as the teacher and `Qwen/Qwen2.5-0.5B` as

Table D.1: **Teacher accuracy (%) across datasets.** Reporting Qwen2.5-1.5B and DeBERTa-v3-base when fine-tuned on full training dataset for each benchmark. These teachers are used as sources of supervision for student models during distillation.

Model	**Amazon Polarity**	**CoLA**	**IMDB**	**SST2**	**Yelp**	**Sentiment140**
Qwen2.5-1.5B	88.5	83.0	94.3	93.7	95.4	86.1
DeBERTa-v3-base	86.7	87.5	93.8	95.8	95.6	86.8

the student, both loaded from Hugging Face with sequence classification heads. We fine-tune using a batch size of 16 and train for 10 epochs. For KD and LWD baselines, we set the distillation loss weights to 20 and 5, respectively. All other settings closely follow the DeBERTaV3 setup, including the optimizer, learning rate schedule, and use of mixed-precision training.

All experiments are conducted on a server equipped with four NVIDIA RTX A6000 GPUs.

Additional Results and Discussion

We provide results using the smaller DeBERTa-v3-xsmall (22M parameters) student as well as the full evaluation table for the Qwen2.5 family. Results for the smaller DeBERTa-v3-xsmall student are shown in Table D.2. While experiments using the Qwen2.5-1.5B teacher and the Qwen2.5-0.5B student are shown in Table D.3. We also include the full fine-tuned teacher model accuracies across all datasets in Table D.1, which are used as supervision targets during knowledge distillation. All experimental results are averaged over five runs, with the mean and standard deviation reported.

Overall, our findings corroborate the central insight that infusing CFEs into knowledge distillation significantly boosts model performance in few-shot settings. For the smaller DeBERTa-v3-xsmall student, we observe that the benefits of CFE infusion remain substantial across tasks, especially when $k \leq 64$. For example, on IMDB at $k = 8$, KD + CoD improves from 74.3% to 89.3%, and LWD

+ CoD improves from 77.3% to 87.7%, showing that even with a much smaller student, CFEs offer a powerful training signal. Similar patterns are seen on SST2 and Amazon Polarity. While the performance gap narrows at higher k values, our method still matches or slightly outperforms standard distillation, despite using only half as many real samples. These results highlight the scalability of CoD across student model sizes.

We also evaluate CoD on Qwen2.5 models, using Qwen2.5-1.5B as the teacher and Qwen2.5-0.5B as the student. Results on CoLA, Yelp, Amazon Polarity, and IMDB show that our method consistently outperforms standard KD and LWD, particularly in few-shot regimes. On IMDB with k=8, KD + CoD reaches 80.0% vs. 67.8% for standard KD - a remarkable 12.2 percentage point gain. Similarly, LWD + CoD improves CoLA accuracy by 8.3 points at k=128 (71.9% vs. 63.6%). With (k=8), CoD boosts Yelp performance by 6.1 points for both KD (74.5% vs. 68.4%) and LWD (74.6% vs. 68.5%). These gains demonstrate the generality of our approach: it is effective even for decoder transformer families like Qwen2.5.

Our findings affirm the broad applicability of CFE-infused distillation. The consistent improvements across datasets, model families, and student capacities support our central claim: CFEs are a powerful, data-efficient tool for improving teacher-student alignment in low-resource scenarios.

Ablation Results

Table D.2: **Classification accuracy (± std) across datasets with varying total training sizes k.** For CoD, training data consists of $k/2$ standard and $k/2$ CFEs. Teacher model DeBERTa-v3-base and student model DeBERTa-v3-xsmall.

		Total Samples (k)					
Dataset	**Method**	**8**	**16**	**32**	**64**	**128**	**512**
Amazon Polarity	KD	0.628 ±0.055	0.690 ±0.034	0.766 ±0.032	0.827 ±0.021	**0.835** ±0.037	0.846 ±0.009
	+ CoD	**0.697** ±0.117	**0.782** ±0.033	**0.823** ±0.018	**0.844** ±0.009	0.814 ±0.013	**0.855** ±0.018
	LWD	0.660 ±0.061	0.699 ±0.044	0.777 ±0.042	0.825 ±0.015	**0.839** ±0.015	0.839 ±0.013
	+ CoD	**0.712** ±0.039	**0.743** ±0.051	**0.811** ±0.016	**0.832** ±0.015	0.830 ±0.010	**0.850** ±0.013
CoLA	KD	0.724 ±0.045	0.735 ±0.052	0.776 ±0.026	0.773 ±0.040	0.799 ±0.011	0.806 ±0.004
	+ CoD	**0.752** ±0.042	**0.766** ±0.018	**0.790** ±0.012	**0.799** ±0.004	**0.803** ±0.008	**0.817** ±0.007
	LWD	**0.699** ±0.042	0.744 ±0.039	0.755 ±0.043	0.787 ±0.008	**0.803** ±0.009	0.808 ±0.008
	+ CoD	0.685 ±0.190	**0.780** ±0.018	**0.790** ±0.004	**0.798** ±0.007	0.802 ±0.005	**0.813** ±0.003
IMDB	KD	0.743 ±0.070	0.849 ±0.037	0.882 ±0.032	**0.904** ±0.004	**0.912** ±0.005	**0.920** ±0.004
	+ CoD	**0.893** ±0.007	**0.896** ±0.007	**0.900** ±0.005	0.904 ±0.005	0.910 ±0.008	0.918 ±0.003
	LWD	0.773 ±0.034	0.823 ±0.041	0.876 ±0.027	**0.903** ±0.008	**0.915** ±0.007	0.914 ±0.014
	+ CoD	**0.877** ±0.022	**0.888** ±0.006	**0.900** ±0.005	0.902 ±0.009	0.911 ±0.008	**0.921** ±0.001
SST2	KD	0.591 ±0.040	0.666 ±0.030	0.754 ±0.047	0.816 ±0.024	0.861 ±0.015	0.887 ±0.033
	+ CoD	**0.685** ±0.112	**0.763** ±0.084	**0.829** ±0.028	**0.850** ±0.015	**0.862** ±0.016	**0.905** ±0.011
	LWD	0.580 ±0.064	0.664 ±0.024	0.726 ±0.036	0.818 ±0.019	0.847 ±0.029	**0.912** ±0.005
	+ CoD	**0.658** ±0.107	**0.675** ±0.074	**0.839** ±0.017	**0.841** ±0.019	**0.859** ±0.016	0.877 ±0.044
Yelp	KD	0.704 ±0.062	**0.793** ±0.042	0.861 ±0.011	0.887 ±0.004	**0.907** ±0.007	**0.922** ±0.008
	+ CoD	**0.759** ±0.086	0.758 ±0.084	**0.870** ±0.008	**0.889** ±0.009	0.897 ±0.009	0.920 ±0.006
	LWD	0.714 ±0.049	**0.815** ±0.028	0.870 ±0.013	0.875 ±0.012	**0.907** ±0.006	**0.925** ±0.006
	+ CoD	**0.758** ±0.069	0.757 ±0.082	**0.873** ±0.012	**0.884** ±0.007	0.894 ±0.009	0.919 ±0.006
Sent140	KD	**0.580** ±0.032	0.594 ±0.026	0.634 ±0.047	0.681 ±0.046	0.740 ±0.012	**0.796** ±0.013
	+ CoD	0.573 ±0.078	**0.612** ±0.064	**0.721** ±0.019	**0.737** ±0.030	**0.767** ±0.014	0.795 ±0.006
	LWD	**0.576** ±0.038	0.585 ±0.025	0.624 ±0.029	0.684 ±0.044	0.728 ±0.035	**0.799** ±0.007
	+ CoD	0.561 ±0.064	**0.592** ±0.050	**0.681** ±0.043	**0.723** ±0.025	**0.763** ±0.019	0.773 ±0.026

Table D.3: **Classification accuracy (± std) of Qwen2.5 across datasets with varying training sizes k.** For CoD, training data consists of $k/2$ standard and $k/2$ CFEs. Teacher model is Qwen2.5-1.5B and student model is Qwen2.5-0.5B.

		Total Samples (k)					
Dataset	**Method**	**8**	**16**	**32**	**64**	**128**	**512**
CoLA	KD	0.681 ±0.012	0.676 ±0.023	0.668 ±0.042	0.654 ±0.032	0.676 ±0.020	0.732 ±0.014
	+ CoD	**0.683** ±0.016	**0.686** ±0.018	**0.697** ±0.015	**0.711** ±0.020	**0.736** ±0.017	**0.757** ±0.011
	LWD	0.681 ±0.012	0.657 ±0.031	0.678 ±0.018	0.650 ±0.039	0.636 ±0.029	0.712 ±0.014
	+ CoD	**0.682** ±0.018	**0.687** ±0.013	**0.704** ±0.010	**0.714** ±0.020	**0.719** ±0.022	**0.755** ±0.013
Yelp	KD	0.684 ±0.021	0.759 ±0.040	0.827 ±0.030	0.861 ±0.017	**0.887** ±0.012	**0.920** ±0.010
	+ CoD	**0.745** ±0.029	**0.779** ±0.048	**0.828** ±0.072	**0.886** ±0.007	0.883 ±0.010	0.916 ±0.008
	LWD	0.685 ±0.019	0.777 ±0.036	0.837 ±0.027	0.876 ±0.020	**0.898** ±0.008	**0.920** ±0.005
	+ CoD	**0.746** ±0.028	**0.778** ±0.035	**0.847** ±0.020	**0.876** ±0.014	0.883 ±0.010	0.909 ±0.009
Amazon Polarity	KD	0.589 ±0.057	0.635 ±0.044	0.706 ±0.083	0.781 ±0.033	**0.807** ±0.031	**0.862** ±0.013
	+ CoD	**0.605** ±0.051	**0.660** ±0.042	**0.712** ±0.077	**0.793** ±0.030	0.805 ±0.041	0.835 ±0.021
	LWD	0.589 ±0.057	0.628 ±0.096	0.680 ±0.052	0.779 ±0.026	0.823 ±0.027	**0.858** ±0.015
	+ CoD	**0.607** ±0.051	**0.662** ±0.060	**0.692** ±0.080	**0.795** ±0.041	**0.823** ±0.023	0.853 ±0.020
IMDB	KD	0.678 ±0.054	0.758 ±0.079	0.817 ±0.057	**0.890** ±0.017	0.903 ±0.012	**0.926** ±0.003
	+ CoD	**0.800** ±0.054	**0.845** ±0.061	**0.877** ±0.038	0.889 ±0.014	**0.912** ±0.010	0.921 ±0.003
	LWD	0.678 ±0.054	0.740 ±0.076	0.832 ±0.035	0.883 ±0.014	0.906 ±0.014	**0.925** ±0.003
	+ CoD	**0.800** ±0.055	**0.835** ±0.048	**0.869** ±0.012	**0.893** ±0.013	**0.909** ±0.008	0.920 ±0.007
SST2	KD	0.568 ±0.061	0.621 ±0.084	0.719 ±0.102	**0.827** ±0.038	**0.878** ±0.020	**0.904** ±0.010
	+ CoD	**0.578** ±0.064	**0.663** ±0.081	**0.767** ±0.085	0.779 ±0.137	0.870 ±0.019	0.886 ±0.005
	LWD	0.568 ±0.062	0.642 ±0.107	0.704 ±0.065	**0.825** ±0.034	**0.869** ±0.026	**0.890** ±0.010
	+ CoD	**0.577** ±0.063	**0.677** ±0.076	**0.782** ±0.133	0.779 ±0.085	0.792 ±0.118	0.878 ±0.011
Sent140	KD	**0.586** ±0.047	**0.599** ±0.047	**0.641** ±0.030	0.708 ±0.027	0.756 ±0.020	**0.813** ±0.010
	+ CoD	0.556 ±0.038	0.591 ±0.046	0.616 ±0.055	**0.711** ±0.061	**0.757** ±0.023	0.805 ±0.010
	LWD	**0.587** ±0.051	**0.596** ±0.038	**0.639** ±0.063	**0.718** ±0.038	0.765 ±0.024	0.805 ±0.011
	+ CoD	0.556 ±0.038	0.588 ±0.059	0.621 ±0.051	0.715 ±0.059	**0.765** ±0.012	**0.805** ±0.008

Table D.4: Performance comparison of KD and CoD variants across varying capacities. Mean and standard deviation are computed over CoD variants. This suggests CoD is not overly sensitive to prompt used.

Method	**8**	**16**	**32**	**64**	**128**	**512**
KD	0.617	0.712	0.757	0.820	0.848	0.899
+ CoD (v1)	0.719	0.781	0.821	0.827	0.853	0.892
+ CoD (v2)	0.754	0.789	0.841	0.872	0.890	0.872
+ CoD (v3)	0.738	0.778	0.819	0.835	0.856	0.901
+ CoD (v4)	0.734	0.783	0.830	0.834	0.883	0.888
CoD *(mean)*	0.736	0.783	0.828	0.842	0.870	0.888
(std)	0.012	0.004	0.009	0.018	0.016	0.010

Table D.5: **Compute/energy ablation on SST2 for KD+CoD and LWD+CoD.** Accuracy increases with k, alongside higher runtime and energy. LWD+CoD is consistently costlier due to intermediate representation alignment.

Method	k	**Accuracy**	**Duration (s)**	**CPU (kWh)**	**GPU (kWh)**	**RAM (kWh)**
KD + CoD	8	0.719	478.13	0.01336	0.00966	0.02479
	16	0.781	488.04	0.01341	0.00972	0.02499
	32	0.821	491.14	0.01382	0.01041	0.02547
	64	0.827	547.58	0.01472	0.13620	0.02723
	128	0.853	569.50	0.01639	0.01634	0.02952
	512	0.892	705.21	0.03120	0.03593	0.04536
LWD + CoD	8	0.694	485.12	0.01263	0.01102	0.02514
	16	0.785	496.07	0.01394	0.01158	0.02572
	32	0.832	517.78	0.01472	0.01245	0.02654
	64	0.830	536.01	0.01515	0.01311	0.02775
	128	0.835	668.52	0.01882	0.01670	0.02960
	512	0.880	814.65	0.04621	0.04712	0.04902

Table D.6: **Effect of soft-label calibration on downstream performance (SST2).**

Method (SST2)	**8**	**16**	**32**	**64**	**128**	**512**
KD (no soft label, α=0)	0.553	0.622	0.697	0.712	0.791	0.815
+ CoD (no soft label, α=0)	0.613	0.651	0.701	0.727	0.793	0.792
KD (random soft label)	0.582	0.533	0.543	0.601	0.617	0.649
+ CoD (random soft label)	0.573	0.548	0.552	0.602	0.623	0.632
KD (default)	0.617	0.712	0.757	0.820	0.848	0.899
+ CoD (default)	0.719	0.781	0.821	0.827	0.853	0.892

Appendix E

E.1 Background on Information Theoretic Measures

We outline key information-theoretic measures pertinent to this paper's discussions.

Definition E.1 (Entropy)**.** *Entropy quantifies the uncertainty or unpredictability of a random variable Z. It is mathematically defined by the equation:*

$$H(Z) = -\sum_{z} \Pr(Z = z) \log \Pr(Z = z). \tag{E.1}$$

Definition E.2 (Mutual Information)**.** *Mutual Information, $\mathrm{I}(Z; \hat{Y})$, quantifies the amount of information obtained about random variable Z through $\hat{Y}$. Specifically, it measures the degree of dependence between two variables, Z and $\hat{Y}$, capturing both linear and non-linear dependencies:*

$$\mathrm{I}(Z; \hat{Y}) = \sum_{z,\hat{y}} \Pr(Z = z, \hat{Y} = \hat{y}) \log \frac{\Pr(Z = z, \hat{Y} = \hat{y})}{\Pr(Z = z) \Pr(\hat{Y} = \hat{y})}. \tag{E.2}$$

Definition E.3 (Conditional Mutual Information)**.** *The conditional mutual information, $I(Z; \hat{Y}|S)$, measures the dependency between Z and $\hat{Y}$, conditioned on S:*

$$\mathrm{I}(Z; \hat{Y}|S) = \sum_{s,z,\hat{y}} \Pr(S = s, Z = z, \hat{Y} = \hat{y}) \log \frac{\Pr(Z = z, \hat{Y} = \hat{y}|S = s)}{\Pr(Z = z|S = s) \Pr(\hat{Y} = \hat{y}|S = s)}, \tag{E.3}$$

or alternatively,

$$\mathrm{I}(Z;\hat{Y}|S) = \sum_{s} \Pr(S=s)\mathrm{I}(Z;\hat{Y}|S=s). \tag{E.4}$$

Mutual Information as a Measure of Fairness. Mutual information can be used as a measure of the unfairness or disparity of a model. Mutual Information has been interpreted as the dependence between sensitive attribute Z and model prediction $\hat{Y}$ (captures correlation as well as all non-linear dependencies). Mutual information is zero if and only if Z and $\hat{Y}$ are independent. This means that if the model's predictions are highly correlated with sensitive attributes, that's a sign of unfairness. Mutual information has been explored in fairness in the context of centralized machine learning in [248, 268, 323].

In recent work, [263] provides another interpretation of mutual information $\mathrm{I}(Z;\hat{Y})$ in fairness as the accuracy of predicting Z from $\hat{Y}$ (or the expected probability of error in correctly guessing Z from $\hat{Y}$) from Fano's inequality. Even in information bottleneck literature, mutual information has been interpreted as a measure of how well one random variable predicts (or, aligns with) the other [324].

For local fairness, we are interested in the dependence between model prediction $\hat{Y}$ and sensitive attributes Z at each and every client, i.e., the dependence between $\hat{Y}$ and Z conditioned on the client S. For example, the disparity at client $S=1$ is $\mathrm{I}(Z;\hat{Y}|S=1)$ (the mutual information (dependence) between model prediction and sensitive attribute conditioned on client $S=1$ (considering data at client $S=1$). Our measure for Local Disparity is the conditional mutual information (dependence) between Z and $\hat{Y}$ conditioned on S, denoted as $\mathrm{I}(Z;\hat{Y}|S)$. Local disparity $\mathrm{I}(Z;\hat{Y}|S) = \sum_s p(s)I(Z;\hat{Y}|S=s)$, is an average of the disparity at each client weighted by the $p(s)$, the proportion of data at client $S=s$. The Local Disparity is zero if and only if all client has zero disparity in their local dataset.

E.2 Proofs for Section 6.5

Lemma 6.1 (Relationship between Global Statistical Parity Gap and $\mathrm{I}(Z;\hat{Y})$)**.** *Let* $\Pr(Z{=}0) = \alpha$. *The gap* $SP_{global} = |\Pr(\hat{Y}=1|Z=1) - \Pr(\hat{Y}=1|Z=0)|$ *is bounded by* $\frac{\sqrt{0.5\,\mathrm{I}(Z;\hat{Y})}}{2\alpha(1-\alpha)}$.

Proof. Mutual information can be expressed as KL divergence:

$$\mathrm{I}(Z;\hat{Y}) = D_{KL}\left(\Pr(\hat{Y},Z)||\Pr(\hat{Y})\Pr(Z)\right). \tag{E.5}$$

Using Pinsker's Inequality [325],

$$d_{TV}(Q_1,Q_2) \leq \sqrt{0.5 D_{KL}(Q_1||Q_2)} \tag{E.6}$$

where, $d_{TV}(Q_1,Q_2)$ is the total variation between two probability distributions Q_1, Q_2.

$$\begin{aligned}
&d_{TV}\left(\Pr(\hat{Y},Z),\Pr(\hat{Y})\Pr(Z)\right) = \frac{1}{2}\sum_{\hat{y},z}\left|\Pr(\hat{Y}=\hat{y},Z=z) - \Pr(\hat{Y}=\hat{y})\Pr(Z=z)\right| \\
&= \sum_z \Pr(Z=z)\sum_{\hat{y},z}\frac{1}{2}\left|\Pr(\hat{Y}=\hat{y}|Z=z) - \Pr(\hat{Y}=\hat{y})\right| \\
&= \frac{1}{2}\Pr(Z=1)\left[|\Pr(\hat{Y}=1|Z=1) - \Pr(\hat{Y}=1)| + |\Pr(\hat{Y}=0|Z=1) - \Pr(\hat{Y}=0)|\right] \\
&+ \frac{1}{2}\Pr(Z=0)\left[|\Pr(\hat{Y}=1|Z=0) - \Pr(\hat{Y}=1)| + |\Pr(\hat{Y}=0|Z=0) - \Pr(\hat{Y}=0)|\right] \\
&= \frac{1}{2}\alpha(1-\alpha)|SP1| + \frac{1}{2}\alpha(1-\alpha)|SP0| + \frac{1}{2}\alpha(1-\alpha)|SP1| + \frac{1}{2}\alpha(1-\alpha)|SP0| \\
&= \alpha(1-\alpha)|SP1| + \alpha(1-\alpha)|SP0|
\end{aligned} \tag{E.7}$$

where $\Pr(Z=0) = 1 - \Pr(Z=1) = \alpha$, and

$$SPi = \Pr(\hat{Y} = i|Z = 1) - \Pr(\hat{Y} = i|Z = 0) = \Pr(\hat{Y} = i|Z = 1) - \Pr(\hat{Y} = i).$$

To complete the proof, we show:

$$\begin{aligned} SP1 &= \Pr(\hat{Y} = 1|Z = 1) - \Pr(\hat{Y} = 1) \\ &= \Pr(\hat{Y} = 1|Z = 1) - \Big(1 - \Pr(\hat{Y} = 0)\Big) \\ &= -1 + \Pr(\hat{Y} = 1|Z = 1) + \Pr(\hat{Y} = 0) \\ &= -\Pr(\hat{Y} = 0|Z = 1) + \Pr(\hat{Y} = 0) = -SP0. \end{aligned}$$

Hence, $|SP1| = |SP0|$. From Eq. (E.7) we have: $2\alpha(1-\alpha)|SP_1| \leq \sqrt{0.5\mathrm{I}(Z;\hat{Y})}$.

Remark E.1 (Tightness of Lemma 6.1)**.** *Since our proof exclusively utilizes Pinsker's inequality, their tightness is equivalent. Given* $\mathrm{I}(Z;\hat{Y}) \leq \min\{H(Z), H(\hat{Y})\} \leq H(\hat{Y})$ *and* $H(Y) \leq 1$ *in binary classification. Hence,* $\mathrm{I}(Z;\hat{Y}) \leq 1$ *which is aligned with the known tight regime of Pinsker's inequality (i.e.,* $D_{KL}(P||Q) \leq 1$*) [325]. The inequality is tighter with smaller mutual information* $\mathrm{I}(Z;\hat{Y})$ *values.*

□

Lemma 6.2. $\mathrm{I}(Z;\hat{Y}|S)=0$ *if and only if* $\Pr(\hat{Y}=1|Z=1, S=s)=\Pr(\hat{Y}=1|Z=0, S=s)$ *at all clients s.*

Proof. We aim to establish that $\mathrm{I}(Z;\hat{Y}|S) = 0$ if and only if $\Pr(\hat{Y} = 1|Z = 1, S = s) = \Pr(\hat{Y} = 1|Z = 0, S = s)$ for all clients s. For brevity, we denote $\Pr(Z = z, S = s, Y = y) = p(z, s, y)$.

Forward Direction: Assume $\mathrm{I}(Z;\hat{Y}|S) = 0$.

From Definition E.3, we have:

$$\mathrm{I}(Z;\hat{Y}|S) = \sum_{s,z,\hat{y}} p(s, z, \hat{y}) \log\left(\frac{p(z, \hat{y}|s)}{p(z|s)p(\hat{y}|s)}\right) = 0.$$

This implies that $\log\left(\frac{p(z,\hat{y}|s)}{p(z|s)p(\hat{y}|s)}\right) = 0$ for all $s, z, \hat{y}$, and consequently $\frac{p(z,\hat{y}|s)}{p(z|s)p(\hat{y}|s)} = 1 \; \forall s$.

Observing that $p(z, \hat{y}|s) = p(z|s)p(\hat{y}|z, s)$, we deduce that $\frac{p(z|s)p(\hat{y}|z,s)}{p(z|s)p(\hat{y}|s)} = 1$.

From this, it directly follows that $p(\hat{y}|z, s) = p(\hat{y}|s)$, and thus $\Pr(\hat{Y} = 1|Z = 1, S = s) = \Pr(\hat{Y} = 1|Z = 0, S = s)$.

Reverse Direction: Assume $\Pr(\hat{Y} = 1|Z = 1, S = s) = \Pr(\hat{Y} = 1|Z = 0, S = s)$ for all s.

This implies $p(\hat{y}|s, z) = p(\hat{y}|s)$ for all $s, z, \hat{y}$. Plugging this into the definition of conditional mutual information, we find $\mathrm{I}(Z; \hat{Y}|S) = 0$.

Thus, both directions of the equivalence are proven, concluding the proof. □

Corollary E.1. *The statistical parity at each client s can be expressed as*

$$|SP_s| \leq \frac{\sqrt{0.5\,\mathrm{I}(Z; \hat{Y}|S = s)}}{2\alpha_s(1 - \alpha_s)}$$

where, $\alpha_s = \Pr(Z = 0|S = s) = 1 - \Pr(Z = 1|S = s)$.

Definition E.4 (Difference Between Local and Global Disparity)**.** *The difference between Global and Local Disparity is:* $\mathrm{I}(Z; \hat{Y}) - \mathrm{I}(Z; \hat{Y}|S) = \mathrm{I}(Z; \hat{Y}; S)$. *This term is the "interaction information," which, unlike other mutual-information-based measures, can be positive or negative.*

Interaction information quantifies the redundancy and synergy present in a system. In FL, positive interaction information indicates a system with high levels of redundancy and Global Disparity, while negative interaction information indicates a system with high levels of synergy and Local Disparity. Interaction information can inform the trade-off between Local and Global Disparity.

E.3 Proofs for Section 6.5.1

The Global and Local Disparity in FL can be decomposed into non-negative terms:

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Red}(Z{:}\hat{Y},S). \tag{6.2}$$

$$\mathrm{I}(Z;\hat{Y}|S) = \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Syn}(Z{:}\hat{Y},S). \tag{6.3}$$

Proof. Equation 6.2 follows directly from the PID terms definition.

$$\mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Red}(Z{:}\hat{Y},S) = \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;\hat{Y}|S) + \mathrm{I}(Z;\hat{Y}) - \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;\hat{Y}|S) = \mathrm{I}(Z;\hat{Y}).$$

Equation 6.3 follows from PID terms definition and the chain rule of mutual information.

$$\begin{aligned}\mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Syn}(Z{:}\hat{Y},S) &= \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;\hat{Y}|S) + \mathrm{I}(Z;\hat{Y},S) - \mathrm{I}(Z;S) - \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;\hat{Y}|S)\\ &= \mathrm{I}(Z;S) + \mathrm{I}(Z;\hat{Y}|S) - \mathrm{I}(Z;S)\\ &= \mathrm{I}(Z;\hat{Y}|S).\end{aligned}$$

Now, we prove the non-negativity property of PID decomposition.

$\mathrm{Uni}(Z{:}\hat{Y}|S) = \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;\hat{Y}|S)$ is non-negative since the conditional mutual information is non-negative by definition.

$\mathrm{Syn}(Z{:}\hat{Y},S) = \mathrm{I}(Z;\hat{Y}|S) - \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;\hat{Y}|S) \geq \mathrm{I}(Z;\hat{Y}|S) - \mathrm{I}(Z;\hat{Y}|S) = 0$

The Redundant Disparity:

$$\mathrm{Red}(Z{:}\hat{Y}, S) = \mathrm{I}(Z;\hat{Y}) - \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;\hat{Y}|S) = \max_{Q\in\Delta_p} \mathrm{I}_Q(\hat{Y};Z) - \mathrm{I}_Q(Z;\hat{Y}|S)$$

First equality holds by definition. Second equality holds since marginals on $(\hat{Y}, Z)$ is fixed in Δ_p, hence, $\max_{Q\in\Delta_p} \mathrm{I}_Q(\hat{Y};Z) = \mathrm{I}(\hat{Y};Z)$.

To prove non-negativity of redundant disparity, we construct a distribution Q_0 such that:

$$\Pr_{Q_0}(Z=z,\hat{Y}=y,S=s) = \frac{\Pr(Z=z,\hat{Y}=y)\Pr(Z=z,S=s)}{\Pr(Z=z)}$$

Next, we show $Q_0 \in \Delta_p$. Recall the set Δ_p in Definition 6.1:

$$\Delta_p = \{Q\in\Delta : \Pr_Q(Z=z,\hat{Y}=y) = \Pr(Z=z,\hat{Y}=y), \Pr_Q(Z=z,S=s) = \Pr(Z=z,S=s)\}.$$

$$\begin{aligned}\Pr_{Q_0}(Z=z,\hat{Y}=y) &= \sum_s \Pr_{Q_0}(Z=z,\hat{Y}=y,S=s) = \sum_s \frac{\Pr(Z=z,\hat{Y}=y)}{\Pr(Z=z)}\Pr(Z=z,S=s)\\ &= \frac{\Pr(Z=z,\hat{Y}=y)}{\Pr(Z=z)}\sum_s \Pr(Z=z,S=s) = \Pr(Z=z,\hat{Y}=y).\end{aligned}$$

$$\begin{aligned}\Pr_{Q_0}(Z=z,S=s) &= \sum_{\hat{y}} \Pr_{Q_0}(Z=z,\hat{Y}=y,S=s) = \sum_{\hat{y}} \frac{\Pr(Z=z,\hat{Y}=y)\Pr(Z=z,S=s)}{\Pr(Z=z)}\\ &= \frac{\Pr(Z=z,S=s)}{\Pr(Z=z)}\sum_{\hat{y}} \Pr(Z=z,\hat{Y}=y) = \Pr(Z=z,S=s).\end{aligned}$$

Marginals of Q_0 satisfy conditions on set Δ_p, hence $Q_0 \in \Delta_p$. Also, note that by construction

of Q_0, $\hat{Y}$ and S are independent conditioned on Z, i.e., $\mathrm{I}_{Q_0}(\hat{Y}; S|Z) = 0$. Hence, we have:

$$\begin{aligned}
\mathrm{Red}(Z{:}\hat{Y}, S) &\overset{(a)}{=} \max_{Q \in \Delta_p} \mathrm{I}_Q(Z; \hat{Y}) - \mathrm{I}_Q(Z; \hat{Y}|S) \\
&\overset{(b)}{\geq} \mathrm{I}_{Q_0}(Z; \hat{Y}) - \mathrm{I}_{Q_0}(Z; \hat{Y}|S) \\
&\overset{(c)}{=} H_{Q_0}(Z) + H_{Q_0}(\hat{Y}) - H_{Q_0}(Z, \hat{Y}) - H_{Q_0}(Z|S) - H_{Q_0}(\hat{Y}|S) + H_{Q_0}(Z, \hat{Y}|S) \\
&\overset{(d)}{=} \mathrm{I}_{Q_0}(\hat{Y}; S) - \mathrm{I}_{Q_0}(\hat{Y}; S|Z) \\
&\overset{(e)}{=} \mathrm{I}_{Q_0}(\hat{Y}; S) \overset{(f)}{\geq} 0.
\end{aligned}$$

Here, (*a*) hold from definition of $\mathrm{Red}(Z{:}\hat{Y}, S)$, (*b*) hold since $Q_0 \in \Delta_p$, (*c*)-(*d*) holds from expressing mutual information in terms of entropy, (*e*) hold since $\mathrm{I}_{Q_0}(\hat{Y}; S|Z) = 0$, (*f*) holds from non-negativity property of mutual information.

□

E.4 Proofs for Section 6.5.2

Theorem 6.1 (Impossibility of Using Local Fairness to Attain Global Fairness)**.** *As long as Redundant Disparity* $\mathrm{Red}(Z{:}\hat{Y}, S) > 0$*, the Global Disparity* $\mathrm{I}(Z; \hat{Y}) > 0$ *even if Local Disparity goes to* 0*.*

Proof. For completeness, we have provided a detailed proof that demonstrates the non-negativity property of the terms involved.

$\mathrm{Uni}(Z{:}\hat{Y}|S) = \min_{Q \in \Delta_p} \mathrm{I}_Q(Z; \hat{Y}|S)$ is non-negative since the conditional mutual information is non-negative by definition.

$\mathrm{Syn}(Z{:}\hat{Y}, S) = \mathrm{I}(Z; \hat{Y}|S) - \min_{Q \in \Delta_p} \mathrm{I}_Q(Z; \hat{Y}|S) \geq \mathrm{I}(Z; \hat{Y}|S) - \mathrm{I}(Z; \hat{Y}|S) = 0.$

The Redundant Disparity:

$$\mathrm{Red}(Z{:}\hat{Y}, S) = \mathrm{I}(Z;\hat{Y}) - \min_{Q\in\Delta_p} \mathrm{I}_Q(Z;\hat{Y}|S) = \max_{Q\in\Delta_p} \mathrm{I}_Q(\hat{Y};Z) - \mathrm{I}_Q(Z;\hat{Y}|S)$$

First equality holds by definition. Second equality holds since marginals on $(\hat{Y}, Z)$ is fixed in Δ_p, hence, $\max_{Q\in\Delta_p} \mathrm{I}_Q(\hat{Y};Z) = \mathrm{I}(\hat{Y};Z)$.

To prove non-negativity of redundant disparity, we construct a distribution Q_0 such that:

$$\Pr_{Q_0}(Z=z,\hat{Y}=y,S=s) = \frac{\Pr(Z=z,\hat{Y}=y)\Pr(Z=z,S=s)}{\Pr(Z=z)}$$

Next, we show $Q_0 \in \Delta_p$. Recall the set Δ_p in Definition 6.1:

$$\Delta_p = \{Q\in\Delta : \Pr_Q(Z=z,\hat{Y}=y) = \Pr(Z=z,\hat{Y}=y), \Pr_Q(Z=z,S=s) = \Pr(Z=z,S=s)\}.$$

$$\begin{aligned}\Pr_{Q_0}(Z=z,\hat{Y}=y) &= \sum_s \Pr_{Q_0}(Z=z,\hat{Y}=y,S=s) = \sum_s \frac{\Pr(Z=z,\hat{Y}=y)}{\Pr(Z=z)}\Pr(Z=z,S=s)\\ &= \frac{\Pr(Z=z,\hat{Y}=y)}{\Pr(Z=z)}\sum_s \Pr(Z=z,S=s) = \Pr(Z=z,\hat{Y}=y).\end{aligned}$$

$$\begin{aligned}\Pr_{Q_0}(Z=z,S=s) &= \sum_{\hat{y}} \Pr_{Q_0}(Z=z,\hat{Y}=y,S=s) = \sum_{\hat{y}} \frac{\Pr(Z=z,\hat{Y}=y)\Pr(Z=z,S=s)}{\Pr(Z=z)}\\ &= \frac{\Pr(Z=z,S=s)}{\Pr(Z=z)}\sum_{\hat{y}} \Pr(Z=z,\hat{Y}=y) = \Pr(Z=z,S=s).\end{aligned}$$

Marginals of Q_0 satisfy conditions on set Δ_p, hence $Q_0 \in \Delta_p$. Also, note that by construction of Q_0, $\hat{Y}$ and S are independent conditioned on Z, i.e., $\mathrm{I}_{Q_0}(\hat{Y}; S|Z) = 0$. Hence, we have:

$$
\begin{aligned}
\mathrm{Red}(Z{:}\hat{Y}, S) &\overset{(a)}{=} \max_{Q \in \Delta_p} \mathrm{I}_Q(Z; \hat{Y}) - \mathrm{I}_Q(Z; \hat{Y}|S) \\
&\overset{(b)}{\geq} \mathrm{I}_{Q_0}(Z; \hat{Y}) - \mathrm{I}_{Q_0}(Z; \hat{Y}|S) \\
&\overset{(c)}{=} H_{Q_0}(Z) + H_{Q_0}(\hat{Y}) - H_{Q_0}(Z, \hat{Y}) - H_{Q_0}(Z|S) - H_{Q_0}(\hat{Y}|S) + H_{Q_0}(Z, \hat{Y}|S) \\
&\overset{(d)}{=} \mathrm{I}_{Q_0}(\hat{Y}; S) - \mathrm{I}_{Q_0}(\hat{Y}; S|Z) \\
&\overset{(e)}{=} \mathrm{I}_{Q_0}(\hat{Y}; S) \overset{(f)}{\geq} 0.
\end{aligned}
$$

Here, (*a*) hold from definition of $\mathrm{Red}(Z{:}\hat{Y}, S)$, (*b*) hold since $Q_0 \in \Delta_p$, (*c*)-(*d*) holds from expressing mutual information in terms of entropy, (*e*) hold since $\mathrm{I}_{Q_0}(\hat{Y}; S|Z) = 0$, (*f*) holds from non-negativity property of mutual information.

Hence, from proposition 6.5.1, we prove Theorem 6.1.

As Local Disparity $\mathrm{I}(Z; \hat{Y}|S) \to 0$, then $\mathrm{Uni}(Z{:}\hat{Y}|S) \to 0$ and $\mathrm{Syn}(Z{:}\hat{Y}, S) \to 0$, therefore the Global Disparity $\mathrm{I}(Z; \hat{Y}) \to \mathrm{Red}(Z{:}\hat{Y}, S) \geq 0$. □

Theorem 6.2 (Global Fairness Does Not Imply Local Fairness)**.** *As long as Masked Disparity* $\mathrm{Syn}(Z{:}\hat{Y}, S) > 0$, *local fairness will not be attained even if global fairness is attained.*

Proof. Proof requires the non-negativity property of PID terms (follows similarly from proof of Theorem 6.1). The argument then goes as follows:

As Global Disparity $\mathrm{I}(Z; \hat{Y}) \to 0$, then $\mathrm{Uni}(Z{:}\hat{Y}|S) \to 0$ and $\mathrm{Red}(Z{:}\hat{Y}, S) \to 0$, therefore the Local Disparity $\mathrm{I}(Z; \hat{Y}|S) \to \mathrm{Syn}(Z{:}\hat{Y}, S) \geq 0$. □

Theorem 6.3 (Necessary and Sufficient Condition to Achieve Global Fairness Using Local Fairness)**.**

If Local Disparity $\mathrm{I}(Z;\hat{Y}|S)$ *goes to zero, then Global Disparity* $\mathrm{I}(Z;\hat{Y})$ *also goes to zero, if and only if the Redundant Disparity* $\mathrm{Red}(Z{:}\hat{Y},S){=}0$*. A sufficient condition for* $\mathrm{Red}(Z{:}\hat{Y},S){=}0$ *is* $Z \perp\!\!\!\perp S$*.*

Proof. From the PID of Local and Global Disparity,

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Red}(Z{:}\hat{Y},S),$$

$$\mathrm{I}(Z;\hat{Y}|S) = \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Syn}(Z{:}\hat{Y},S).$$

Therefore if, $\mathrm{I}(Z;\hat{Y}|S) = 0$, then $\mathrm{Uni}(Z{:}\hat{Y}|S) = 0$. Hence,

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Red}(Z{:}\hat{Y},S)$$

$$\mathrm{I}(Z;\hat{Y}) = 0 \iff \mathrm{Red}(Z{:}\hat{Y},S) = 0.$$

To prove the sufficient condition, we leverage the PID of $\mathrm{I}(Z;S)$ and the non-negative property of the PID terms:

$$\mathrm{I}(Z;S) = \mathrm{Uni}(Z{:}S|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y},S)$$

$$\mathrm{I}(Z;S) \geq \mathrm{Red}(Z{:}\hat{Y},S).$$

Hence, $Z \perp\!\!\!\perp S \implies \mathrm{Red}(Z{:}\hat{Y},S) = 0$. □

A sufficient condition for $\mathrm{Red}(Z{:}\hat{Y},S) = 0$ is $\mathrm{Syn}(Z{:}\hat{Y},S) = 0$ and $\hat{Y} \perp\!\!\!\perp S$.

Proof. Interaction information expressed in PID terms (see Definition E.4):

$$\mathrm{I}(Z;\hat{Y};S) = \mathrm{I}(Z;\hat{Y}) - \mathrm{I}(Z;\hat{Y}|S) = \mathrm{Red}(Z{:}\hat{Y},S) - \mathrm{Syn}(Z;\hat{Y},S).$$

If Masked Disparity $\mathrm{Syn}(Z;\hat{Y},S) = 0$, we have:

$$\mathrm{I}(Z;\hat{Y};S) = \mathrm{I}(Z;\hat{Y}) - \mathrm{I}(Z;\hat{Y}|S) = \mathrm{Red}(Z{:}\hat{Y},S) \geq 0$$

Since the interaction information is positive and symmetric,

$$\mathrm{I}(\hat{Y};S) \geq \mathrm{I}(\hat{Y};S) - \mathrm{I}(\hat{Y};S|Z) = \mathrm{Red}(Z{:}\hat{Y},S).$$

Therefore, $\hat{Y} \perp\!\!\!\perp S \implies \mathrm{Red}(Z{:}\hat{Y},S) = 0.$ □

Theorem 6.4. *Local disparity will always be less than Global Disparity if and only if Masked Disparity* $\mathrm{Syn}(Z{:}\hat{Y},S) = 0$. *A sufficient condition is when* $Z - \hat{Y} - S$ *form a Markov chain.*

Proof. By leveraging the PID of $\mathrm{I}(Z;S|\hat{Y})$,

$$\mathrm{I}(Z;S|\hat{Y}) = \mathrm{Uni}(Z{:}S|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},S).$$

Markov chain $Z - \hat{Y} - S$ implies, $\mathrm{I}(Z;S|\hat{Y}) = 0$. Hence, $\mathrm{Syn}(Z{:}\hat{Y},S) = 0$.

Rest of proof follows from nonnegative property of PID terms:

$$\mathrm{I}(Z;\hat{Y}|S) = \mathrm{Uni}(Z{:}\hat{Y}|S) \leq \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Red}(Z{:}\hat{Y},S) = \mathrm{I}(Z;\hat{Y}).$$

□

E.5 Proof of Theorem 6.5

Definition E.5 (Convex Function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if, for all $x_1, x_2 \in \mathbb{R}^n$ and for all $\lambda \in [0, 1]$, the following inequality holds:*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2). \tag{E.8}$$

Lemma E.1 (Log Sum Inequality)**.** *The log-sum inequality states that for any two sequences of non-negative numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, the following inequality holds:*

$$\sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{n} a_i\right) \log\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right). \tag{E.9}$$

Theorem 6.5. *The AGLFOP is a convex optimization problem.*

Proof. The set Δ_p is a convex set, since for any two points $Q_1, Q_2 \in \Delta_p$, their convex combination also lies in Δ_p (probability simplex). To prove AGFOP is a convex optimization problem, we show each term is convex in Q using the definition of a convex function (see Definition E.5).

Let $Q \in \Delta_p$ denote the joint distribution for $(Z, S, Y, \hat{Y})$. For brevity, we denote $\Pr(Z = z, S = s, Y = y) = p(z, s, y)$, the fixed marginals on (Z, S, Y). Additionally, we denote $\Pr_Q(Z = z, S = s, Y = y, \hat{Y} = \hat{y}) = Q(z, s, y, \hat{y})$ as the probability under distribution Q.

To prove that $err(Q)$ is convex, we need to show: $err(Q_\lambda) \leq \lambda err(Q_1) + (1 - \lambda)err(Q_2)$, where $Q_\lambda = \lambda Q_1 + (1 - \lambda)Q_2$, $\forall Q_1, Q_2 \in \Delta_p$, and $\lambda \in [0, 1]$.

We first express $err(Q_\lambda)$ as:

$$\begin{aligned} err(Q_\lambda) &= \sum_{z,s,y,\hat{y}} Q_\lambda(z,s,y,\hat{y}) \cdot \mathbb{I}(y \neq \hat{y}) \\ &= \lambda \sum_{z,s,y,\hat{y}} Q_1(z,s,y,\hat{y}) \cdot \mathbb{I}(y \neq \hat{y}) + (1-\lambda) \sum_{z,s,y,\hat{y}} Q_2(z,s,y,\hat{y}) \cdot \mathbb{I}(y \neq \hat{y}) \\ &= \lambda err(Q_1) + (1-\lambda) err(Q_2). \end{aligned}$$

Thus, $err(Q)$ is convex as it satisfies the convexity condition.

To prove convexity of $\mathrm{I}_Q(Z;\hat{Y})$, we show that $\forall Q_1, Q_2 \in \Delta_p$ and $\forall \lambda \in [0,1]$, the following inequality holds: $I_{Q_\lambda}(Z:\hat{Y}) \leq \lambda I_{Q_1}(Z:\hat{Y}) + (1-\lambda) I_{Q_2}(Z:\hat{Y})$.

Note that the marginals are convex in the joint distribution, i.e., $Q_\lambda(z,\hat{y}) = \sum_{s,y} Q_\lambda(z,s,y,\hat{y})$.

This is not necessarily true for conditionals. However, when *some* marginals are fixed, convexity holds for *some* conditionals, i.e., $Q_\lambda(\hat{y}|z) = \sum_{s,y} Q_\lambda(z,s,y,\hat{y})/p(z)$.

Note that the conditional $Q_\lambda(\hat{y}|z)$ is convex in the joint distribution $Q_\lambda(z,s,y,\hat{y})$.

$$\begin{aligned} Q_\lambda(\hat{y}|z) &= \sum_{s,y} Q_\lambda(z,s,y,\hat{y})/p(z) = \sum_{s,y} Q_\lambda(\hat{y}|z,s,y) p(s,y,z)/p(z) \\ &= \sum_{s,y} Q_\lambda(\hat{y}|z,s,y) p(s,y|z). \end{aligned}$$

Hence, we can show convexity of $\mathrm{I}_Q(Z;\hat{Y})$ in $Q_\lambda(\hat{y}|z)$:

$$\begin{aligned}
\mathrm{I}_{Q_\lambda}(Z;\hat{Y}) &= \sum_{z,\hat{y}} Q_\lambda(z,\hat{y}) \log\left(\frac{Q_\lambda(z,\hat{y})}{Q_\lambda(z)Q_\lambda(\hat{y})}\right) \\
&= \sum_{z,\hat{y}} Q_\lambda(z)Q_\lambda(\hat{y}|z) \log\left(\frac{Q_\lambda(\hat{y}|z)}{Q_\lambda(\hat{y})}\right) \\
&\overset{(a)}{=} \sum_{z,\hat{y}} p(z)(\lambda Q_1(\hat{y}|z) + (1-\lambda)Q_2(\hat{y}|z)) \log\left(\frac{\lambda Q_1(\hat{y}|z) + (1-\lambda)Q_2(\hat{y}|z)}{\lambda Q_1(\hat{y}) + (1-\lambda)Q_2(\hat{y})}\right) \\
&\overset{(b)}{\leq} \lambda \sum_{z,\hat{y}} p(z)Q_1(\hat{y}|z) \log\left(\frac{Q_1(\hat{y}|z)}{Q_1(\hat{y})}\right) + (1-\lambda) \sum_{z,\hat{y}} p(z)Q_2(\hat{y}|z) \log\left(\frac{Q_2(\hat{y}|z)}{Q_2(\hat{y})}\right) \\
&= \lambda \mathrm{I}_{Q_1}(Z;\hat{Y}) + (1-\lambda)\mathrm{I}_{Q_2}(Z;\hat{Y}).
\end{aligned}$$

Here (a) holds from expressing the linear combinations. Also note that, $Q_\lambda(\hat{y}) = \sum_z Q_\lambda(\hat{y}|z)p(z)$, which can also be expressed as a linear combination. The inequality (b) holds from the log-sum inequality (see Lemma E.1).

To prove the convexity of $\mathrm{I}_Q(Z;\hat{Y}|S)$, we show that $\forall Q_1, Q_2 \in \Delta_p$ and $\forall \lambda \in [0,1]$, the following inequality holds: $\mathrm{I}_{Q_\lambda}(Z;\hat{Y}|S) \leq \lambda \mathrm{I}_{Q_1}(Z;\hat{Y}|S) + (1-\lambda)\mathrm{I}_{Q_2}(Z;\hat{Y}|S)$.

Note that the conditional $Q_\lambda(\hat{y}|z,s)$ is convex in the joint distribution $Q_\lambda(z,s,y,\hat{y})$:

$$\begin{aligned}
Q_\lambda(\hat{y}|z,s) &= \sum_y Q_\lambda(\hat{y},z,s,y)/p(z,s) = \sum_y Q_\lambda(\hat{y}|z,s,y)p(y,z,s)/p(z,s) \\
&= \sum_y Q_\lambda(\hat{y}|z,s,y)p(y|z,s).
\end{aligned}$$

Hence, we can show convexity of $\mathrm{I}_Q(Z;\hat{Y}|S)$ in $Q_\lambda(\hat{y}|z,s)$:

$$
\begin{aligned}
I_{Q_\lambda}(Z;\hat{Y}|S) &= \sum_{z,s,\hat{y}} Q_\lambda(z,s,\hat{y}) \log\left(\frac{Q_\lambda(\hat{y}|z,s)}{Q_\lambda(\hat{y}|s)}\right) \\
&\overset{(a)}{=} \sum_{z,s,\hat{y}} p(z,s)\left(\lambda Q_1(\hat{y}|z,s) + (1-\lambda)Q_2(\hat{y}|z,s)\right) \log\left(\frac{\lambda Q_1(\hat{y}|z,s) + (1-\lambda)Q_2(\hat{y}|z,s)}{\lambda Q_1(\hat{y}|s) + (1-\lambda)Q_2(\hat{y}|s)}\right) \\
&\overset{(b)}{\leq} \lambda \sum_{z,s,\hat{y}} p(z,s) Q_1(\hat{y}|z,s) \log\left(\frac{Q_1(\hat{y}|z,s)}{Q_1(\hat{y}|s)}\right) + (1-\lambda) \sum_{z,s,\hat{y}} p(z,s) Q_2(\hat{y}|z,s) \log\left(\frac{Q_2(\hat{y}|z,s)}{Q_2(\hat{y}|s)}\right) \\
&= \lambda \mathrm{I}_{Q_1}(Z;\hat{Y}|S) + (1-\lambda)\mathrm{I}_{Q_2}(Z;\hat{Y}|S).
\end{aligned}
$$

The equality (a) holds from linear combinations of $Q_\lambda(\hat{y}|s) = \sum_z Q_\lambda(\hat{y}|z,s)p(z|s)$. The inequality (b) holds due to the application of the log-sum inequality (see Lemma E.1). □

E.6 Expanded Experiments

This section includes additional results, expanded tables, figures, and details that provide a more comprehensive understanding of our study.

Dataset. We consider the following datasets:

(1) Synthetic dataset: A 2-D feature vector $X = (X_0, X_1)$ follows a distribution given by $X|_{Y=1} \sim \mathcal{N}((2,2), \left[\begin{smallmatrix} 5 & 1 \\ 1 & 5 \end{smallmatrix}\right])$, $X|_{Y=0} \sim \mathcal{N}((-2,-2), \left[\begin{smallmatrix} 10 & 1 \\ 1 & 3 \end{smallmatrix}\right])$. Assume Z is a binary sensitive attribute such that $Z = 1$ if $X_0 > 0$, else $Z = 0$, to encode dependence of X_0 with Z.

(2) Adult dataset: The Adult dataset is a publicly available dataset in the UCI repository based on 1994 U.S. census data [37]. The goal is to predict whether an individual earns more or less than $50,000 per year based on features such as occupation, marital status, and education.

Client Distribution. We strategically partition our datasets across clients to examine scenarios char-

acterized by Unique, Redundant, and Masked Disparities.

Scenario 1: Uniform Distribution of Sensitive Attributes Across Clients. The sensitive attribute Z is independently distributed across clients, i.e., $Z \perp\!\!\!\perp S$. We randomly distribute the data across clients.

Scenario 2: High Heterogeneity in Sensitive Attributes Across Clients. We split to observe heterogeneity in the distribution of sensitive attributes across clients, i.e., $Z = S$ with a probability α. For instance, when $\alpha = 0.9$, the client with $S = 0$ consists of 90% people with $Z = 0$, while the client with $S = 1$ is composed of 90% people with $Z = 1$. For the Adult dataset, we use $\alpha = \Pr(Z = 0|S = 0)$ as a parameter to regulate this heterogeneity.

Scenario 3: High Synergy Level Across Clients. The true label $Y \approx Z \oplus S$. To emulate this scenario, we partition the data such that client $S = 0$ possesses data of people with $Z = 1$ and true labels $Y = 1$ and people with $Z = 0$ and true labels $Y = 0$. Conversely, client $S = 1$ contains the remaining data, i.e., $Z = 1$ with $Y = 0$ and $Z = 0$ with $Y = 1$.

We introduce the *synergy level* (Definition E.6) to measure alignment to $Y = Z \oplus S$.

Definition E.6 (Synergy Level (λ))**.** *The synergy level $\lambda \in [0, 1]$ of a given dataset and client distribution is defined as the probability that the true label Y is aligned with $Z \oplus S$,*

$$\lambda = \Pr(Y = Z \oplus S),$$

where $\lambda = 1$ implies perfect alignment between Y and $Z \oplus S$, and $\lambda = 0$ implies zero alignment.

To set λ when splitting data across clients, we first split with perfect XOR alignment and then shuffle fractions of the dataset between clients.

Adult Heterogeneous Split. In Fig. 6.4 *(fourth row)*, we split the Adult dataset to capture various disparities simultaneously. We set the synergy level $\lambda = 0.8$ (see Definition E.6). Due to the nature

of the Adult dataset, this introduces some correlation between the sensitive attribute Z and client S.

Experiment A:Accuracy-Global-Local-Fairness Trade-off Pareto Front.

To study the trade-offs between model accuracy and different fairness constraints, we plot the Pareto frontiers for the AGLFOP. We solve for maximum accuracy $(1 - err)$ while varying global and local fairness relaxations (ϵ_g, ϵ_l). We present results for synthetic and Adult datasets as well as PID terms for various data splitting scenarios across clients. The three-way trade-off among accuracy, global, and local fairness can be visualized as a contour plot (see Fig. 6.4).

For the Adult dataset, we restrict our optimization space Δ_p to lie within the convex hull derived by the False Positive Rate (FPR) and True Positive Rate (TPR) of an initially trained classifier (trained using FedAvg). This characterizes the accuracy-fairness for all derived classifiers from the original trained classifier (motivated by the post-processing technique from [246]). The convex hall characterizes the distributions that can be achieved with any derived classifier. The convex hull for each protected group is composed of points, including $(0, 0)$, (TPR, FPR), $(\overline{TPR}, \overline{FPR})$ and $(1, 1)$, where $\overline{TPR}$ and $\overline{FPR}$ denote the true positive and false positive rates of a predictor that inverts all predictions for a protected group. Future work could explore alternative constraints for various specialized applications.

Experiment B: Demonstrating Disparities in Federated Learning Settings.

In this experiment, we investigate the PID of disparities in the Adult dataset trained within a FL framework. We employ the *FedAvg* algorithm [283] for training.

Setup. Our FL model employs a two-layer architecture with 32 hidden units per layer, using ReLU

activation, binary cross-entropy loss, and the Adam optimizer. The server initializes the model weights and distributes them to clients, who train locally on partitioned datasets for 2 epochs with a batch size of 64. Client-trained weights are aggregated server-side via the FedAvg algorithm, and this process iterates until convergence. Evaluation metrics are estimated using the `dit` package [287], which includes PID functions for decomposing Global and Local Disparities into Unique, Redundant, and Masked Disparity, following the definition from [284].

We analyze the following scenarios:

PID of Disparity Across Various Splitting Scenarios. We partition the dataset across two clients, each time varying the level of sensitive attribute heterogeneity ($\alpha = \Pr(Z = 0|S = 0)$). The Adult dataset exhibits a sensitive attribute imbalance with $\Pr(Z = 0) = 0.33$, making $\alpha = \Pr(Z = 0|S = 0) = 0.33$ the independently distributed case.

In this first setup, we distribute the *sensitive attribute uniformly across clients (splitting scenario 1)* and employ FedAvg for training. The FL model achieves an accuracy of 84.45% with a Global Disparity of 0.0359 bits and a Local Disparity of 0.0359 bits. The PID reveals that the Unique Disparity is 0.0359 bits, with both Redundant and Masked Disparities being negligible. This aligns with our centralized baseline, indicating that the disparity originates exclusively from the dependency between the model's predictions and the sensitive attributes, rather than being influenced by S.

When the dataset is split to introduce *high heterogeneity in sensitive attributes across clients (splitting scenario* 2), the resulting FL model exhibits a Global Disparity of 0.0431 bits and a Local Disparity of 0.0014 bits. PID reveals a Redundant Disparity of 0.0431 bits and a Masked Disparity of 0.0014 bits, with no Unique Disparity.

Next we split and train according to *splitting scenario* 3 *($\lambda = 0.9$)*. The trained model reports a Local Disparity of 0.1761 bits and a Global Disparity of 0.0317 bits. The PID decomposition shows a

Masked Disparity of 0.1761 bits and a Redundant Disparity of 0.0317 bits, with no Unique Disparity observed. The emergence of non-zero Redundant Disparity is attributable to the data splitting, which consequently leads to $\mathrm{I}(Z;S) = 0.2409$ bits.

We summarize the three scenarios in Fig. 6.5. Additionally, we evaluate the effects of using a naive local disparity mitigation technique on the various disparities present.

Effects of Naive Local Fairness Mitigation Technique. We evaluate the effects of using a naive local disparity mitigation technique on the various disparities present. This is achieved by incorporating a statistical parity regularizer to the loss function at each individual client:

$$\texttt{client\_loss} = \texttt{client\_cross\_entropy\_loss} + \beta\ \texttt{client\_fairness\_loss}.$$

We use an implementation from FairTorch package [311]. The term β is a hyperparameter that trades off between accuracy and fairness. We use $\beta = 0.1$ to maintain similarly accurate models. The results are presented in Table E.1.

Table E.1: Table illustrates the effects of using a naive local disparity mitigation technique on the various scenarios. It proved efficacious only when Unique Disparity is present (*scenario 1*). However, with high redundancy or synergy (*scenarios 2 &3*), the utilization of the disparity mitigation technique exacerbated disparities.

	Loc.	Glob.	Uniq.	Red.	Mas.
Scenario 1	0.0359	0.0359	0.0359	0.0000	0.0000
+ fairness	0.0062	0.0062	0.0062	0.0000	0.0000
Scenario 2	0.0014	0.0431	0.0000	0.0431	0.0014
+ fairness	0.0110	0.0626	0.0000	0.0626	0.0110
Scenario 3	0.1761	0.0317	0.0000	0.0317	0.1761
+ fairness	0.0935	0.0418	0.0053	0.0365	0.0882

PID of Disparity under Heterogeneous Sensitive Attribute Distribution. We analyze the PID of Local and Global Disparities under different sensitive attribute distributions across clients. We train the model with two clients, each having equal-sized datasets. We use $\alpha = \Pr(Z = 0|S = 0)$ to represent

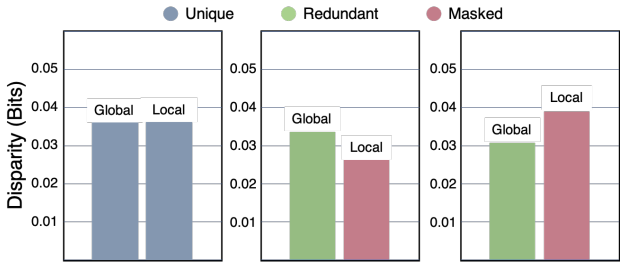


Figure E.1: Plot demonstrating scenarios with Unique, Redundant, and Masked Disparities for the Adult dataset 5 client case. Difficulty in splitting to achieve pure Redundant and Masked Disparity due to the proportion of labels in the dataset.

sensitive attribute heterogeneity. Note that for a fixed α, the proportions of sensitive attributes at the other client are fixed. For example since $\Pr(Z = 0) = 0.33$ for the Adult dataset, $\alpha = 0.33$ results in even distribution of sensitive attributes across the two clients. Our results are summarized in Fig. 6.5 and Table E.2. We also provide results for 10 federating clients in Table E.4.

Table E.2: The PID of Global and Local Disparity for varying sensitive attribute heterogeneity α

α	$\mathrm{I}(Z;S)$	Local	Global	Unique	Redundant	Masked	$\mathrm{I}(\hat{Y};S)$	Accuracy
0.1	0.1877	0.0262	0.0342	0.0000	0.0342	0.0262	0.0080	86.54%
0.2	0.0575	0.0336	0.0364	0.0064	0.0301	0.0273	0.0028	86.95%
0.3	0.0032	0.0363	0.0365	0.0332	0.0032	0.0031	0.0002	86.86%
0.33	0.0000	0.0340	0.0340	0.0340	0.0000	0.0000	0.0000	87.34%
0.4	0.0154	0.0311	0.0319	0.0186	0.0133	0.0125	0.0009	86.70%
0.5	0.0957	0.0368	0.0413	0.0023	0.0390	0.0345	0.0045	86.77%
0.6	0.2613	0.0242	0.0346	0.0000	0.0346	0.0242	0.0104	86.61%
0.66	0.4392	0.0185	0.0325	0.0000	0.0325	0.0185	0.0140	86.36%

Observing Levels of Masked Disparity. We aim to gain a deeper understanding of the circumstances with Masked Disparities. Through scenario 3, we showed how high Masked Disparities can occur. However, the level of synergy portrayed in the example may not always be present in practice. We attempt to quantify this using a metric *synergy level*. The synergy level (λ) measures how closely the true labels Y align with the XOR of Z and S (see Definition E.6). To achieve a high synergy level,

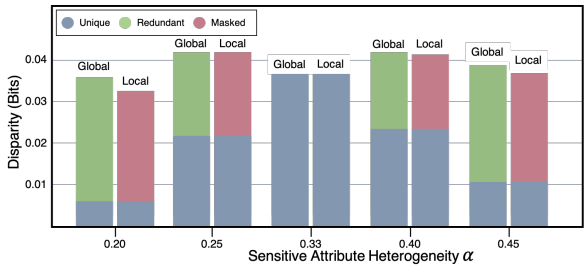


Figure E.2: Plot showing the PID of disparities when the data is near i.i.d. among $K = 10$ clients. All types of disparities can be observed. The value $\alpha = 0.33$ represents the case where the data is i.i.d. and only Unique Disparity is observed.

we apply the method outlined in Scenario 3. To decrease λ, we randomly shuffle data points between clients until the synergy level reaches 0. We conduct experiments with varying levels of synergy to observe the impact on the Masked Disparity. The results are summarized in Fig. 6.5 and Table E.3.

Table E.3: PID of Global and Local Disparity under varying synergy levels λ

λ	$I(Z;S)$	Loc.	Glob.	Uniq.	Red.	Mas.	$\mathrm{I}(\hat{Y};S)$	Acc.	$\mathrm{I}(Z;\hat{Y}\|S=0)$	$\mathrm{I}(Z;\hat{Y}\|S=1)$
0	0.0035	0.0402	0.0373	0.0338	0.0035	0.0063	0.0005	85.12%	0.0196	0.0608
0.25	0.0113	0.0486	0.0419	0.0308	0.0111	0.0178	0.0009	85.54%	0.0819	0.0152
0.5	0.0299	0.0536	0.0335	0.0127	0.0208	0.0410	0.0033	85.24%	0.1056	0.0017
0.75	0.0846	0.0932	0.0366	0.0023	0.0343	0.0909	0.0068	85.26%	0.0024	0.1840
1	0.2409	0.1644	0.0149	0.0000	0.0150	0.1644	0.0201	84.30%	0.0839	0.2450

Multiple Client Case. We examine scenarios involving multiple clients. Observations are similar to the two-client case previously studied. To observe a high Unique Disparity, sensitive attributes need to be identically distributed across clients. To observe the Redundant Disparity, there must be some dependency between clients and a specific sensitive attribute, meaning certain demographic groups are known to belong to a specific client. The Masked Disparity can be observed when there is a high level of synergy or XOR behavior between variables Z and S. Note that since S is no longer binary, we can convert its decimal value to binary and then take the XOR.

We experiment with $K = 5$ clients and examine the three disparities. To observe the Unique

Disparity by randomly distributing the data among clients. For Redundant Disparity, we divide the data such that the first two clients are mostly $Z = 0$ and the remaining three clients are mostly $Z = 1$. For Masked Disparity, we distribute the data similarly to scenario 3 (see Fig. E.1 and Fig. E.2).

Additional Insights from Experiments. When data is uniformly distributed across clients, Unique Disparity is dominant and contributes to both global and local unfairness (see Fig. 6.5 *Scenario 1*: model trained using FedAvg on the adult dataset and distributed uniformly across clients). In the trade-off Pareto Front (see Fig. 6.4, *row 1*), we see that both local and global fairness constraints have balanced tradeoffs with accuracy. The PID decomposition (Fig. 6.5, *row 1*, *column 2,3,4*) explains this as we see the disparity is mainly Unique Disparity, with little Redundant or Masked Disparity. The Unique Disparity highlights where Local and Global Disparity agree.

Case with sensitive attribute heterogeneity (sensitive attribute imbalance across clients): We observe mainly Redundant Disparity (see Fig. 6.5, *scenario 2* and *middle*), this is a globally unfair but locally fair model (recall Proposition 6.5.1). Observe in the tradeoff plot (see Fig. 6.4, *row 2*) that the accuracy trade-off is with mainly global fairness (an accurate model could have zero Local Disparity but be globally unfair).

Cases with sensitive-attribute synergy across clients: In a two-client case (one client is more likely to have qualified candidates with $Z = 0$ and unqualified $Z = 1$ and vice versa at the other client). We observe that the Mask Disparity is dominant (see Fig. 6.5, *Scenario 3*). The trade AGLFOP tradeoff plot (see Fig. 6.4, *row 3*) is characterized by Masked Disparity with trade-offs mainly between local fairness and accuracy (an accurate model could have zero Global Disparity but be locally unfair). The Redundant and Masked Disparity highlights where Local and Global Disparity disagree.

The AGLFOP provides the theoretical boundaries trade-offs, capturing the optimal performance any model or FL technique can achieve for a specified dataset and client distribution. For example, say

Table E.4: PID of Global & Local Disparity for various sensitive attribute distributions across 10 clients.

α	Unique	Redundant	Masked	Global	Local	Accuracy
0.25	0.0219	0.0190	0.0178	0.0409	0.0409	84.85%
0.33	0.0376	0.0000	0.0000	0.0376	0.0376	85.58%
0.4	0.0268	0.0141	0.0137	0.0410	0.0405	84.85%
0.45	0.0107	0.0289	0.0270	0.0390	0.0377	84.85%

one wants a perfectly globally fair and locally fair model, i.e., ($\epsilon_g = 0$, $\epsilon_l = 0$). Under high sensitive attribute heterogeneity (see Fig. 6.4, *row 2*), they cannot have a model that does better than 64%.

Appendix F

F.1 Proofs for Section 7.4.1

Theorem 7.1 (Revisiting Impossibility)**.** *If* $\mathrm{I}(Z;\hat{Y},Y)>0$*, at least one of the PID terms, namely,* $\mathrm{Uni}(Z{:}\hat{Y}|Y)$*,* $\mathrm{Red}(Z{:}\hat{Y},Y)$*,* $\mathrm{Syn}(Z{:}\hat{Y},Y)$*, or* $\mathrm{Uni}(Z{:}Y|\hat{Y})$ *will be nonnegative. Hence, at least one of the fairness measures, namely, the Statistical Parity Gap (*$\mathrm{I}(Z;\hat{Y})$*), Equalized Odds Gap (*$\mathrm{I}(Z;\hat{Y}|Y)$*), or Predictive Parity Gap (*$\mathrm{I}(Z;Y|\hat{Y})$*) will be nonzero. Conversely, all these unfairness measures will be zero if and only if* $\mathrm{I}(Z;\hat{Y},Y)=0$*.*

Proof. For completeness, we first show the non-negativity of PID terms for the PID definition that we are using in this work (also see [284]):

$\mathrm{Uni}(Z{:}\hat{Y}|Y)=\min_{Q\in\Delta_p}\mathrm{I}_Q(Z;\hat{Y}|Y)$ is non-negative since the conditional mutual information is non-negative by definition.

Similarly argument holds for $\mathrm{Uni}(Z{:}Y|\hat{Y})$.

$\mathrm{Syn}(Z{:}\hat{Y},Y)=\mathrm{I}(Z;\hat{Y}|Y)-\min_{Q\in\Delta_p}\mathrm{I}_Q(Z;\hat{Y}|Y)\geq\mathrm{I}(Z;\hat{Y}|Y)-\mathrm{I}(Z;\hat{Y}|Y)=0.$

The Redundant information:

$$\mathrm{Red}(Z{:}\hat{Y},Y)=\mathrm{I}(Z;\hat{Y})-\min_{Q\in\Delta_p}\mathrm{I}_Q(Z;\hat{Y}|S)=\max_{Q\in\Delta_p}\mathrm{I}_Q(\hat{Y};Z)-\mathrm{I}_Q(Z;\hat{Y}|Y)$$

First equality holds by definition of redundant information. Second equality holds since marginals on

$(\hat{Y}, Z)$ is fixed in Δ_p, hence, $\max_{Q \in \Delta_p} \mathrm{I}_Q(\hat{Y}; Z) = \mathrm{I}(\hat{Y}; Z)$.

To prove non-negativity of redundant information, we construct a distribution Q_0 such that:

$$\Pr_{Q_0}(Z = z, \hat{Y} = y, Y = y) = \frac{\Pr(Z = z, \hat{Y} = y) \Pr(Z = z, Y = y)}{\Pr(Z = z)}$$

Next, we show $Q_0 \in \Delta_p$. Recall the set Δ_p in Definition 6.1:

$$\Delta_p = \{Q \in \Delta : \Pr_Q(Z = z, \hat{Y} = y) = \Pr(Z = z, \hat{Y} = y), \Pr_Q(Z = z, Y = y) = \Pr(Z = z, Y = y)\}.$$

$$\begin{aligned}\Pr_{Q_0}(Z = z, \hat{Y} = y) &= \sum_y \Pr_{Q_0}(Z = z, \hat{Y} = y, Y = y) = \sum_y \frac{\Pr(Z = z, \hat{Y} = y)}{\Pr(Z = z)} \Pr(Z = z, Y = y) \\ &= \frac{\Pr(Z = z, \hat{Y} = y)}{\Pr(Z = z)} \sum_y \Pr(Z = z, Y = y) = \Pr(Z = z, \hat{Y} = y).\end{aligned}$$

$$\begin{aligned}\Pr_{Q_0}(Z = z, Y = y) &= \sum_{\hat{y}} \Pr_{Q_0}(Z = z, \hat{Y} = y, Y = y) = \sum_{\hat{y}} \frac{\Pr(Z = z, \hat{Y} = y) \Pr(Z = z, Y = y)}{\Pr(Z = z)} \\ &= \frac{\Pr(Z = z, Y = y)}{\Pr(Z = z)} \sum_{\hat{y}} \Pr(Z = z, \hat{Y} = y) = \Pr(Z = z, Y = y).\end{aligned}$$

Marginals of Q_0 satisfy conditions on set Δ_p, hence $Q_0 \in \Delta_p$. Also, note that by construction

of Q_0, $\hat{Y}$ and Y are independent conditioned on Z, i.e., $\mathrm{I}_{Q_0}(\hat{Y};Y|Z)=0$. Hence, we have:

$$\begin{aligned}
\mathrm{Red}(Z{:}\hat{Y},Y) &\overset{(a)}{=} \max_{Q\in\Delta_p} \mathrm{I}_Q(Z;\hat{Y}) - \mathrm{I}_Q(Z;\hat{Y}|Y) \\
&\overset{(b)}{\geq} \mathrm{I}_{Q_0}(Z;\hat{Y}) - \mathrm{I}_{Q_0}(Z;\hat{Y}|Y) \\
&\overset{(c)}{=} H_{Q_0}(Z) + H_{Q_0}(\hat{Y}) - H_{Q_0}(Z,\hat{Y}) - H_{Q_0}(Z|Y) - H_{Q_0}(\hat{Y}|Y) + H_{Q_0}(Z,\hat{Y}|Y) \\
&\overset{(d)}{=} \mathrm{I}_{Q_0}(\hat{Y};Y) - \mathrm{I}_{Q_0}(\hat{Y};Y|Z) \\
&\overset{(e)}{=} \mathrm{I}_{Q_0}(\hat{Y};Y) \overset{(f)}{\geq} 0.
\end{aligned}$$

Here, (*a*) hold from definition of $\mathrm{Red}(Z{:}\hat{Y},Y)$, (*b*) hold since $Q_0\in\Delta_p$, (*c*)-(*d*) holds from expressing mutual information in terms of entropy, (*e*) hold since $\mathrm{I}_{Q_0}(\hat{Y};S|Z)=0$, (*f*) holds from non-negativity property of mutual information.

Since mutual information $\mathrm{I}(Z;\hat{Y},Y)>0$, we can use the PID framework to decompose this mutual information as:

$$\mathrm{I}(Z;\hat{Y},Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y},Y) + \mathrm{Syn}(Z{:}\hat{Y},Y).$$

Since each of the PID components $\mathrm{Uni}(Z{:}\hat{Y}|Y)$, $\mathrm{Uni}(Z{:}Y|\hat{Y})$, $\mathrm{Red}(Z{:}\hat{Y},Y)$, $\mathrm{Syn}(Z{:}\hat{Y},Y)$ is non-negative, the condition $\mathrm{I}(Z;\hat{Y},Y)>0$ implies that at least one of these terms must be strictly positive.

Next, according to Proposition 7.1:

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y},Y),$$

$$\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Syn}(Z{:}\hat{Y},Y),$$

$$\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y).$$

Given that $\mathrm{I}(Z;\hat{Y},Y) > 0$ ensures that at least one of $\mathrm{Uni}(Z{:}\hat{Y}|Y)$, $\mathrm{Red}(Z{:}\hat{Y},Y)$, $\mathrm{Syn}(Z{:}\hat{Y},Y)$, or $\mathrm{Uni}(Z{:}Y|\hat{Y})$ is nonnegative, it follows that:

If $\mathrm{Uni}(Z{:}\hat{Y}|Y) > 0$, then $\mathrm{I}(Z;\hat{Y}) > 0$.

If $\mathrm{Syn}(Z{:}\hat{Y},Y) > 0$, then $\mathrm{I}(Z;\hat{Y}|Y) > 0$ or $\mathrm{I}(Z;Y|\hat{Y}) > 0$.

If $\mathrm{Red}(Z{:}\hat{Y},Y) > 0$, then $\mathrm{I}(Z;\hat{Y}) > 0$.

If $\mathrm{Uni}(Z{:}Y|\hat{Y}) > 0$, then $\mathrm{I}(Z;Y|\hat{Y}) > 0$.

Therefore, the presence of positive mutual information $\mathrm{I}(Z;\hat{Y},Y)$ guarantees that at least one of the fairness measures ($\mathrm{I}(Z;\hat{Y})$, $\mathrm{I}(Z;\hat{Y}|Y)$, $\mathrm{I}(Z;Y|\hat{Y})$) will be nonzero. Conversely, if all these unfairness measures are zero, then by the definitions given in Proposition 7.1 and non-negativity of PID terms, all the PID terms must be zero, which implies $\mathrm{I}(Z;\hat{Y},Y) = 0$. □

Theorem 7.2 (Dataset Dependent Relationships)**.** *If* $\mathrm{I}(Z;Y)>0$*, either the Statistical Parity Gap* $\mathrm{I}(Z;\hat{Y})$ *or the Predictive Parity Gap* $\mathrm{I}(Z;Y|\hat{Y})$ *must be greater than zero.*

Proof. We begin by expressing the mutual information between Z and Y using PID:

$$\mathrm{I}(Z;Y) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}Y,\hat{Y}). \tag{F.1}$$

Next, we examine the contributions of these PID terms to the fairness measures: The term

$\mathrm{Uni}(Z{:}Y|\hat{Y})$ contributes to the Predictive Parity Gap ($\mathrm{I}(Z;Y|\hat{Y})$). The term $\mathrm{Red}(Z{:}Y,\hat{Y})$ contributes to the Statistical Parity Gap ($\mathrm{I}(Z;\hat{Y})$).

Given that $\mathrm{I}(Z;Y) > 0$, it follows that the sum of $\mathrm{Uni}(Z{:}Y|\hat{Y})$ and $\mathrm{Red}(Z{:}Y,\hat{Y})$ is positive, i.e., $\mathrm{I}(Z;Y) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}Y,\hat{Y}) > 0$.

Since both $\mathrm{Uni}(Z{:}Y|\hat{Y})$ and $\mathrm{Red}(Z{:}Y,\hat{Y})$ are nonnegative, the fact that their sum is positive implies that at least one of them must be strictly positive. Thus, we have two cases to consider:

1. If $\mathrm{Uni}(Z{:}Y|\hat{Y}) > 0$, then the Predictive Parity Gap ($\mathrm{I}(Z;Y|\hat{Y})$) must be greater than zero.

2. If $\mathrm{Red}(Z{:}Y,\hat{Y}) > 0$, then the Statistical Parity Gap ($\mathrm{I}(Z;\hat{Y})$) must be greater than zero.

Therefore, if $\mathrm{I}(Z;Y) > 0$, it is guaranteed that either $\mathrm{I}(Z;\hat{Y}) > 0$ or $\mathrm{I}(Z;Y|\hat{Y}) > 0$. □

F.2 Proofs for Section 7.4.2

Theorem 7.3. *If Statistical Parity is satisfied, i.e.,* $\mathrm{I}(Z;\hat{Y}) = 0$*, then the Predictive Parity Gap is greater than the Equalized Odds Gap, i.e.,* $\mathrm{I}(Z;Y|\hat{Y}) \geq \mathrm{I}(Z;\hat{Y}|Y)$*. Additionally, if the dataset is such that* $\mathrm{I}(Z;Y) = 0$*, then Predictive Parity and Equalized Odds are equivalent, i.e.,* $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{I}(Z;\hat{Y}|Y)$.

Proof. Given that Statistical Parity is zero ($\mathrm{I}(Z;\hat{Y}) = 0$), we have:

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y},Y) = 0. \tag{F.2}$$

Since all PID terms are non-negative, this implies:

$$\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0 \quad \text{and} \quad \mathrm{Red}(Z{:}\hat{Y},Y) = 0. \tag{F.3}$$

The Equalized Odds gap ($\mathrm{I}(Z;\hat{Y}|Y)$) and Predictive Parity gap ($\mathrm{I}(Z;Y|\hat{Y})$) can be expressed as:

$$\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{Syn}(Z{:}\hat{Y},Y), \tag{F.4}$$

$$\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y). \tag{F.5}$$

Since $\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0$, we have:

$$\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y) \geq \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{I}(Z;\hat{Y}|Y). \tag{F.6}$$

Furthermore, if $\mathrm{I}(Z;Y) = 0$, we have:

$$\mathrm{I}(Z;Y) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y},Y) = 0. \tag{F.7}$$

This implies:

$$\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0 \quad \text{and} \quad \mathrm{Red}(Z{:}\hat{Y},Y) = 0. \tag{F.8}$$

Therefore:

$$\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{I}(Z;\hat{Y}|Y). \tag{F.9}$$

Thus, we have: $\mathrm{I}(Z;Y|\hat{Y}) \geq \mathrm{I}(Z;\hat{Y}|Y)$ when $\mathrm{I}(Z;\hat{Y}) = 0$, and $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{I}(Z;\hat{Y}|Y)$ when $\mathrm{I}(Z;Y) = 0$. □

Theorem 7.4. *If Predictive Parity is satisfied, i.e.,* $\mathrm{I}(Z;Y|\hat{Y})=0$*, then the Statistical Parity Gap is*

greater than the Equalized Odds Gap, i.e., $\mathrm{I}(Z;\hat{Y}) \geq \mathrm{I}(Z;\hat{Y}|Y)$*. Additionally, if the dataset is such that* $\mathrm{I}(Z;Y) = 0$*, then Statistical Parity and Equalized Odds are equal, i.e.,* $\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{I}(Z;\hat{Y}|Y)$.

Proof. Given that Predictive Parity is satisfied ($\mathrm{I}(Z;Y|\hat{Y}) = 0$), we have:

$$\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y},Y) = 0. \tag{F.10}$$

Since all PID terms are non-negative, this implies:

$$\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0 \quad \text{and} \quad \mathrm{Syn}(Z{:}\hat{Y},Y) = 0. \tag{F.11}$$

The Statistical Parity gap ($\mathrm{I}(Z;\hat{Y})$) and Equalized Odds gap ($\mathrm{I}(Z;\hat{Y}|Y)$) can be expressed as:

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y},Y), \tag{F.12}$$

$$\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Syn}(Z{:}\hat{Y},Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y). \tag{F.13}$$

Since $\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0$, we have:

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y},Y) \geq \mathrm{Uni}(Z{:}\hat{Y}|Y) = \mathrm{I}(Z;\hat{Y}|Y). \tag{F.14}$$

Thus, $\mathrm{I}(Z;\hat{Y}) \geq \mathrm{I}(Z;\hat{Y}|Y)$ when $\mathrm{I}(Z;Y|\hat{Y}) = 0$.

Furthermore, if $\mathrm{I}(Z;Y) = 0$, we have:

$$\mathrm{I}(Z;Y) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y},Y) = 0. \tag{F.15}$$

This implies:

$$\mathrm{Uni}(Z{:}Y|\hat{Y}) = 0 \quad \text{and} \quad \mathrm{Red}(Z{:}\hat{Y}, Y) = 0. \tag{F.16}$$

Therefore: $\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) \quad \text{and} \quad \mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y)$.

Hence, when $\mathrm{I}(Z;Y) = 0$, and $\mathrm{I}(Z;Y|\hat{Y}) = 0$, we have: $\mathrm{I}(Z;\hat{Y}) = \mathrm{I}(Z;\hat{Y}|Y)$. □

Theorem 7.5. *If Equalized Odds is satisfied, i.e.,* $\mathrm{I}(Z;\hat{Y}|Y)=0$ *and* $\mathrm{I}(Z;Y) > 0$*, an inverse relationship (tradeoff) exists between Statistical Parity and Predictive Parity, i.e.,* $\mathrm{I}(Z;\hat{Y}) = \mathrm{I}(Z;Y) - \mathrm{I}(Z;Y|\hat{Y})$*. Thus, increasing one leads to a decrease in the other, and vice versa.*

Proof. Given that Equalized Odds is satisfied ($\mathrm{I}(Z;\hat{Y}|Y) = 0$), we have:

$$\mathrm{I}(Z;\hat{Y}|Y) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Syn}(Z{:}\hat{Y}, Y) = 0. \tag{F.17}$$

Since all PID terms are non-negative, it follows that:

$$\mathrm{Uni}(Z{:}\hat{Y}|Y) = 0 \quad \text{and} \quad \mathrm{Syn}(Z{:}\hat{Y}, Y) = 0. \tag{F.18}$$

The Statistical Parity gap ($\mathrm{I}(Z;\hat{Y})$) and Predictive Parity gap ($\mathrm{I}(Z;Y|\hat{Y})$) can be expressed as:

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|Y) + \mathrm{Red}(Z{:}\hat{Y}, Y) = \mathrm{Red}(Z{:}\hat{Y}, Y), \tag{F.19}$$

$$\mathrm{I}(Z;Y|\hat{Y}) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Syn}(Z{:}\hat{Y}, Y) = \mathrm{Uni}(Z{:}Y|\hat{Y}). \tag{F.20}$$

Given that $\mathrm{I}(Z;Y) > 0$, we can write the mutual information between Z and Y as:

$$\mathrm{I}(Z;Y) = \mathrm{Uni}(Z{:}Y|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y},Y). \tag{F.21}$$

Substituting $\mathrm{I}(Z;Y|\hat{Y})$ and $\mathrm{I}(Z;\hat{Y})$ into this equation, we get:

$$\mathrm{I}(Z;Y) = \mathrm{I}(Z;Y|\hat{Y}) + \mathrm{I}(Z;\hat{Y}). \tag{F.22}$$

Since $\mathrm{I}(Z;Y)$ is fixed for a given dataset, this equation demonstrates that $\mathrm{I}(Z;\hat{Y})$ and $\mathrm{I}(Z;Y|\hat{Y})$ have an inverse relationship:

$$\mathrm{I}(Z;\hat{Y}) = \mathrm{I}(Z;Y) - \mathrm{I}(Z;Y|\hat{Y}). \tag{F.23}$$

Therefore, increasing the Statistical Parity gap ($\mathrm{I}(Z;\hat{Y})$) will lead to a decrease in the Predictive Parity gap ($\mathrm{I}(Z;Y|\hat{Y})$), and vice versa. □

Bibliography

[1] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023.

[2] Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784, 2022.

[3] R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1998.

[4] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), feb 2022.

[5] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning, 2019.

[7] Kush R Varshney. *Trustworthy Machine Learning*. Independently Published, Chappaqua, NY, 2021.

[8] Consumer Financial Protection Bureau. 12 cfr part 1002 - equal credit opportunity act (regulation b). Accessed: 2024-11-06.

[9] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[10] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[11] A Saranya and R Subhashini. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, 7:100230, 2023.

[12] L. Breiman. Statistical modeling: The two cultures. *Quality Engineering*, 48:81–82, 2001.

[13] Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR, 2020.

[14] 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*, FAccT '22, page 850–863, 2022.

[15] Hsiang Hsu and Flavio Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. In *Advances in Neural Information Processing Systems*, volume 35, pages 28988–29000. Curran Associates, Inc., 2022.

[16] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *International Conference on Machine Learning*, pages 12351–12367. PMLR, 2023.

[17] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust algorithmic recourse under model multiplicity with probabilistic guarantees. *IEEE Journal on Selected Areas in Information Theory*, 2024.

[18] Faisal Hamman, Pasan Dissanayake, Saumitra Mishra, Freddy Lecue, and Sanghamitra Dutta. Quantifying prediction consistency under fine-tuning multiplicity in tabular LLMs. In *Forty-second International Conference on Machine Learning*, 2025.

[19] Faisal Hamman, Chenyang Zhu, Anoop Kumar, Xujun Peng, Sanghamitra Dutta, Daben Liu, and Alfy Samuel. Improving consistency in retrieval-augmented systems with group similarity reward. In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*, 2025.

[20] Faisal Hamman, Pasan Dissanayake, Yanjun Fu, and Sanghamitra Dutta. Few-shot knowledge distillation of llms with counterfactual explanations. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

[21] Faisal Hamman and Sanghamitra Dutta. Demystifying local and global fairness trade-offs in federated learning using partial information decomposition. *International Conference on Learning Representations (ICLR)*, 2024.

[22] Sanghamitra Dutta and Faisal Hamman. A review of partial information decomposition in algorithmic fairness and explainability. *Entropy*, 25(5):795, 2023.

[23] Faisal Hamman and Sanghamitra Dutta. A unified view of group fairness tradeoffs using partial information decomposition. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 214–219, 2024.

[24] Faisal Hamman, Jiahao Chen, and Sanghamitra Dutta. Can Querying for Bias Leak Protected Attributes? Achieving Privacy With Smooth Sensitivity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1358–1368, 2023.

[25] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[26] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *CoRR*, abs/2010.04050, 2020.

[27] Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.

[28] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. *Advances in Neural Information Processing Systems*, 34, 2021.

[29] Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. Consistent counterfactuals for deep models. *arXiv preprint arXiv:2110.03109*, 2021.

[30] Satyapriya Krishna, Jiaqi Ma, and Himabindu Lakkaraju. Towards bridging the gaps between the right to explanation and the right to be forgotten. *arXiv preprint arXiv:2302.04288*, 2023.

[31] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pages 3126–3132, 2020.

[32] Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. Robust counterfactual explanations for tree-based ensembles. In *International Conference on Machine Learning*, pages 5742–5756. PMLR, 2022.

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

[34] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

[35] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Robust counterfactual explanations in machine learning: A survey. *arXiv preprint arXiv:2402.01928*, 2024.

[36] FICO. FICO XML Challenge. `https://community.fico.com/s/explainable-machine-learning-challenge`, 2018.

[37] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[38] I-Cheng Yeh and Che hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2, Part 1):2473–2480, 2009.

[39] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pages 809–818. PMLR, 2020.

[40] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In *IJCAI*, pages 2855–2862, 2020.

[41] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[42] Leif Hancox-Li. Robustness in Machine Learning Explanations: Does It Matter? In *Proceedings of the 3rd ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 640–647. Barcelona, Spain, January 27–30 2020.

[43] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Can I still trust you?: Understanding the impact of distribution shifts on algorithmic recourses. *arXiv preprint arXiv:2012.11788*, 2020.

[44] Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. Provably robust and plausible counterfactual explanations for neural networks via robust optimisation. *arXiv preprint arXiv:2309.12545*, 2023.

[45] Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. A Survey on the Robustness of Feature Importance and Counterfactual Explanations. *arXiv e-prints*, arXiv:2111.00358, 2021.

[46] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Issues with post-hoc counterfactual explanations: a discussion. *arXiv preprint arXiv:1906.04774*, 2019.

[47] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[48] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically robust recourse: Navigating the trade-offs between costs and robustness in algorithmic recourse. *arXiv preprint arXiv:2203.06768*, 2022.

[49] Donato Maragno, Jannis Kurtz, Tabea E Röber, Rob Goedhart, Ş Ilker Birbil, and Dick den Hertog. Finding regions of counterfactual explanations via robust optimization. *arXiv preprint arXiv:2301.11113*, 2023.

[50] Ricardo Dominguez-Olmedo, Amir H Karimi, and Bernhard Schölkopf. On the adversarial robustness of causal algorithmic recourse. In *International Conference on Machine Learning*, pages 5324–5342. PMLR, 2022.

[51] Jamelle Watson-Daniels, David C Parkes, and Berk Ustun. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10306–10314, 2023.

[52] Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. Counterfactual shapley additive explanations. *ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[53] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[54] Hsueh-Ti Derek Liu, Francis Williams, Alec Jacobson, Sanja Fidler, and Or Litany. Learning smooth neural functions via lipschitz regularization. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–13, 2022.

[55] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive approximation*, 28:253–263, 2008.

[56] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[57] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[58] Scikit-Learn. LOF Implementation.

[59] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.

[60] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

[61] Dan Ley, Saumitra Mishra, and Daniele Magazzeni. Global counterfactual explanations: Investigations, implementations and improvements, 2022.

[62] Natraj Raman, Daniele Magazzeni, and Sameena Shah. Bayesian hierarchical models for counterfactual estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 1115–1128. PMLR, 2023.

[63] Ahmad-Reza Ehyaei, Amir-Hossein Karimi, Bernhard Schölkopf, and Setareh Maghsudi. Robustness implies fairness in casual algorithmic recourse. *arXiv preprint arXiv:2302.03465*, 2023.

[64] Boris van Breugel and Mihaela van der Schaar. Why tabular foundation models should be a research priority. *arXiv preprint arXiv:2405.01147*, 2024.

[65] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*, 2020.

[66] Jiahuan Yan, Bo Zheng, Hongxia Xu, Yiheng Zhu, Danny Chen, Jimeng Sun, Jian Wu, and Jintai Chen. Making pre-trained language models great on tabular prediction. *arXiv preprint arXiv:2403.01841*, 2024.

[67] Ruiyu Wang, Zifeng Wang, and Jimeng Sun. Unipredict: Large language models are universal tabular predictors. *arXiv preprint arXiv:2310.03266*, 2023.

[68] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[69] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[70] Leo Breiman. Statistical modeling: The two cultures. *Quality control and applied statistics*, 48(1):81–82, 2003.

[71] Juan Felipe Gomez, Caio Machado, Lucas Monteiro Paes, and Flavio Calmon. Algorithmic arbitrariness in content moderation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2234–2253, 2024.

[72] Yuwei Yin, Yazheng Yang, Jian Yang, and Qi Liu. Finpt: Financial risk prediction with profile tuning on pretrained foundation models. *arXiv preprint arXiv:2308.00065*, 2023.

[73] Zifeng Wang, Chufan Gao, Cao Xiao, and Jimeng Sun. Meditab: Scaling medical tabular data predictors via data consolidation, enrichment, and refinement, 2024.

[74] Zekai Chen, Mariann Micsinai Balan, and Kevin Brown. Language models are few-shot learners for prognostic prediction. *arXiv preprint arXiv:2302.12692*, 2023.

[75] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*, 2024.

[76] Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey, 2024.

[77] 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.

[78] Kacper Sokol, Meelis Kull, Jeffrey Chan, and Flora Dilys Salim. Cross-model fairness: Empirical study of fairness and ethics under model multiplicity, 2023.

[79] 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.

[80] Michael Kahn. Diabetes. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5T59G.

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

[82] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

[83] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[84] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. Deep entity matching with pre-trained language models. *arXiv preprint arXiv:2004.00584*, 2020.

[85] Avanika Narayan, Ines Chami, Laurel Orr, Simran Arora, and Christopher Ré. Can foundation models wrangle your data? *arXiv preprint arXiv:2205.09911*, 2022.

[86] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

[87] Dimitris Bertsimas, Kimberly Villalobos Carballo, Yu Ma, Liangyuan Na, Léonard Boussioux, Cynthia Zeng, Luis R Soenksen, and Ignacio Fuentes. Tabtext: a systematic approach to aggregate knowledge across tabular data structures. *arXiv preprint arXiv:2206.10381*, 2022.

[88] Soma Onishi, Kenta Oono, and Kohei Hayashi. Tabret: Pre-training transformer-based tabular models for unseen columns. *arXiv preprint arXiv:2303.15747*, 2023.

[89] Han Zhang, Xumeng Wen, Shun Zheng, Wei Xu, and Jiang Bian. Towards foundation models for learning on tabular data. *arXiv preprint arXiv:2310.07338*, 2023.

[90] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654, 2024.

[91] Yazheng Yang, Yuqi Wang, Sankalok Sen, Lei Li, and Qi Liu. Unleashing the potential of large language models for predictive tabular tasks in data science. *arXiv preprint arXiv:2403.20208*, 2024.

[92] Xiangyang Li, Bo Chen, Lu Hou, and Ruiming Tang. Ctrl: Connect tabular and language model for ctr prediction. *arXiv preprint arXiv:2306.02841*, 2023.

[93] Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. Large language models (llms) on tabular data: Predic-tion, generation, and understanding-a survey. *arXiv preprint arXiv:2402.17944*, 2024.

[94] Kathleen Creel and Deborah Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1):26–43, 2022.

[95] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems*, 35:14071–14084, 2022.

[96] Kota Mata, Kentaro Kanamori, and Hiroki Arimura. Computing the collection of good models for rule lists. *arXiv preprint arXiv:2204.11285*, 2022.

[97] Hsiang Hsu, Ivan Brugere, Shubham Sharma, Freddy Lecue, and Chun-Fu Chen. Rashomongb: Analyzing the rashomon effect and mitigating predictive multiplicity in gradient boosting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[98] 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.

[99] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2144–2155. PMLR, 18–24 Jul 2021.

[100] Kacper Sokol, Meelis Kull, Jeffrey Chan, and Flora Dilys Salim. Fairness and ethics under model multiplicity in machine learning. *arXiv preprint arXiv:2203.07139*, 2022.

[101] Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. Predictive multiplicity in probabilistic classification, 2023.

[102] 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.

[103] Josip Djolonga, Frances Hubis, Matthias Minderer, Zachary Nado, Jeremy Nixon, Rob Romijnders, Dustin Tran, and Mario Lucic. Robustness Metrics. `https://github.com/google-research/robustness_metrics`, 2020. Version 0.0.1.

[104] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Efficient estimation of the local robustness of machine learning models. *arXiv preprint arXiv:2307.13885*, 2023.

[105] Sukriti Jaitly, Tanay Shah, Ashish Shugani, and Razik Singh Grewal. Towards better serialization of tabular data for few-shot classification. *arXiv preprint arXiv:2312.12464*, 2023.

[106] Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. *arXiv preprint arXiv:2310.17513*, 2023.

[107] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[108] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[109] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995.

[110] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 09–15 Jun 2019.

[111] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in neural information processing systems*, 32, 2019.

[112] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[113] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics.

[114] Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems*, 2023.

[115] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[116] Vinay Chamola, Vikas Hassija, A Razia Sulthana, Debshishu Ghosh, Divyansh Dhingra, and Biplab Sikdar. A review of trustworthy and explainable artificial intelligence (xai). *IEEE Access*, 2023.

[117] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253):1–15, 2023.

[118] Spurthi Amba Hombaiah, Tao Chen, Mingyang Zhang, Michael Bendersky, and Marc Najork. Dynamic language models for continuously evolving content. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2514–2524, 2021.

[119] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[120] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

[121] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

[122] Jekaterina Novikova, Carol Anderson, Borhane Blili-Hamelin, and Subhabrata Majumdar. Consistency in language models: Current landscape, challenges, and future directions. *arXiv preprint arXiv:2505.00268*, 2025.

[123] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

[124] Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. Improving consistency in large language models through chain of guidance. *arXiv preprint arXiv:2502.15924*, 2025.

[125] Sezen Perçin, Xin Su, Qutub Sha Syed, Phillip Howard, Aleksei Kuvshinov, Leo Schwinn, and Kay-Ulrich Scholl. Investigating the robustness of retrieval-augmented generation at the query level. *arXiv preprint arXiv:2507.06956*, 2025.

[126] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

[127] Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. *arXiv preprint arXiv:2508.21038*, 2025.

[128] Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. Benchmarking prompt sensitivity in large language models. In *European Conference on Information Retrieval*, pages 303–313. Springer, 2025.

[129] Sunnie SY Kim, Jennifer Wortman Vaughan, Q Vera Liao, Tania Lombrozo, and Olga Russakovsky. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2025.

[130] Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. Becel: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3680–3696, 2022.

[131] Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. A logic-driven framework for consistency of neural models. *arXiv preprint arXiv:1909.00126*, 2019.

[132] Akari Asai and Hannaneh Hajishirzi. Logic-guided data augmentation and regularization for consistent question answering. *arXiv preprint arXiv:2004.10157*, 2020.

[133] Eric Mitchell, Joseph J Noh, Siyan Li, William S Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D Manning. Enhancing self-consistency and performance of pretrained language models through natural language inference. *arXiv preprint arXiv:2211.11875*, 2022.

[134] Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*, 2020.

[135] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

[136] Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. Evaluating the factual consistency of large language models through news summarization. *arXiv preprint arXiv:2211.08412*, 2022.

[137] Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. *arXiv preprint arXiv:2311.07466*, 2023.

[138] Vamshi Krishna Bonagiri, Sreeram Vennam, Priyanshul Govil, Ponnurangam Kumaraguru, and Manas Gaur. Sage: Evaluating moral consistency in large language models. *arXiv preprint arXiv:2402.13709*, 2024.

[139] Alexios Arvanitis and Konstantinos Kalliris. Consistency and moral integrity: A self-determination theory perspective. *Journal of Moral Education*, 49(3):316–329, 2020.

[140] Faisal Hamman, Pasan Dissanayake, Saumitra Mishra, Freddy Lecue, and Sanghamitra Dutta. Quantifying prediction consistency under fine-tuning multiplicity in tabular LLMs. In *Forty-second International Conference on Machine Learning*, 2025.

[141] Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. Measuring reliability of large language models through semantic consistency. *arXiv preprint arXiv:2211.05853*, 2022.

[142] Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby-Tavor. Predicting question-answering performance of large language models through semantic consistency. *arXiv preprint arXiv:2311.01152*, 2023.

[143] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

[144] Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Shuaiqiang Wang, Chong Meng, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. Improving the robustness of large language models via consistency alignment. *arXiv preprint arXiv:2403.14221*, 2024.

[145] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[146] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.

[147] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[148] Jennifer Hsia, Afreen Shaikh, Zora Zhiruo Wang, and Graham Neubig. Ragged: Towards informed design of scalable and stable rag systems. In *Forty-second International Conference on Machine Learning*, 2025.

[149] Kepu Zhang, Zhongxiang Sun, Weijie Yu, Xiaoxue Zang, Kai Zheng, Yang Song, Han Li, and Jun Xu. Qe-rag: A robust retrieval-augmented generation benchmark for query entry errors. *arXiv preprint arXiv:2504.04062*, 2025.

[150] Zhibo Hu, Chen Wang, Yanfeng Shu, Hye-Young Paik, and Liming Zhu. Prompt perturbation in retrieval-augmented generation based large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1119–1130, 2024.

[151] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback, 2024.

[152] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[153] John C Gower and Pierre Legendre. Metric and euclidean properties of dissimilarity coefficients. *Journal of classification*, 3(1):5–48, 1986.

[154] Yapei Chang, Yekyung Kim, Michael Krumdick, Amir Zadeh, Chuan Li, Chris Tanner, and Mohit Iyyer. Bleuberi: Bleu is a surprisingly effective reward for instruction following. *arXiv preprint arXiv:2505.11080*, 2025.

[155] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

[156] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[157] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*, 2020.

[158] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

[159] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. In *Proceedings of ACL 2019*, 2019.

[160] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[161] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rockt"aschel, and Sebastian Riedel. KILT: a Benchmark for Knowledge Intensive Language Tasks. *CoRR*, 2020.

[162] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

[163] Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. Inference scaling for long-context retrieval augmented generation. *arXiv preprint arXiv:2410.04343*, 2024.

[164] Yuetong Zhao, Hongyu Cao, Xianyu Zhao, and Zhijian Ou. An empirical study of retrieval augmented generation with chain-of-thought. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 436–440. IEEE, 2024.

[165] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

[166] OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2024.

[167] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and et al. Qwen2.5 technical report. *arXiv:2412.15115*, 2025.

[168] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*, 2025.

[169] Sanjay Surendranath Girija, Shashank Kapoor, Lakshit Arora, Dipen Pradhan, Aman Raj, and Ankit Shetgaonkar. Optimizing llms for resource-constrained environments: A survey of model compression techniques. *arXiv preprint arXiv:2505.02309*, 2025.

[170] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[171] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.

[172] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[173] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

[174] Sangchul Hahn and Heeyoul Choi. Self-knowledge distillation in natural language processing. *arXiv preprint arXiv:1908.01851*, 2019.

[175] Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

[176] Yuanhao Yue, Chengyu Wang, Jun Huang, and Peng Wang. Distilling instruction-following abilities of large language models with task-aware curriculum planning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6030–6054, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[177] Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*, pages 20852–20867. PMLR, 2023.

[178] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

[179] Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Haoran Lu, Xilin Gong, Yufang Liu, et al. Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *arXiv preprint arXiv:2504.14772*, 2025.

[180] Hui Xue, Yuexuan An, Yongchun Qin, Wenqian Li, Yixin Wu, Yongjuan Che, Pengfei Fang, and Minling Zhang. Towards few-shot learning in the open world: A review and beyond. *arXiv preprint arXiv:2408.09722*, 2024.

[181] Min Zhang, Donglin Wang, and Sibo Gai. Knowledge distillation for model-agnostic meta-learning. In *ECAI 2020*, pages 1355–1362. IOS Press, 2020.

[182] Benlin Liu, Yongming Rao, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Metadistiller: Network self-boosting via meta-learned top-down distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 694–709. Springer, 2020.

[183] Tianhong Li, Jianguo Li, Zhuang Liu, and Changshui Zhang. Few sample knowledge distillation for efficient network compression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14639–14647, 2020.

[184] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017.

[185] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[186] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[187] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[188] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review, 2022.

[189] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions, 2020.

[190] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

[191] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

[192] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019.

[193] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[194] Pasan Dissanayake, Faisal Hamman, Barproda Halder, Ilia Sucholutsky, Qiuyi Zhang, and Sanghamitra Dutta. Quantifying knowledge distillation using partial information decomposition. *arXiv preprint arXiv:2411.07483*, 2024.

[195] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[196] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[197] Jiaheng Liu, Chenchen Zhang, Jinyang Guo, Yuanxing Zhang, Haoran Que, Ken Deng, Jie Liu, Ge Zhang, Yanan Wu, Congnan Liu, et al. DDK: Distilling domain knowledge for efficient large language models. *Advances in Neural Information Processing Systems*, 37:98297–98319, 2024.

[198] Chengliang Chai, Jiabin Liu, Nan Tang, Ju Fan, Dongjing Miao, Jiayi Wang, Yuyu Luo, and Guoliang Li. Goodcore: Data-effective and data-efficient machine learning through coreset selection over incomplete data. *Proceedings of the ACM on Management of Data*, 1(2):1–27, 2023.

[199] Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. How to train data-efficient llms. *arXiv preprint arXiv:2402.09668*, 2024.

[200] Suruchi Kumari and Pravendra Singh. Data efficient deep learning for medical image analysis: A survey. *arXiv preprint arXiv:2310.06557*, 2023.

[201] Pasan Dissanayake and Sanghamitra Dutta. Model reconstruction using counterfactual explanations: A perspective from polytope theory. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[202] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput. Surv.*, 55(5), December 2022.

[203] Volker Kaibel and Marc E Pfetsch. Some algorithmic problems in polytope theory. In *Algebra, geometry and software systems*, pages 23–47. Springer, 2003.

[204] Van Bach Nguyen, Paul Youssef, Christin Seifert, and Jörg Schlötterer. Llms for generating and evaluating counterfactuals: A comprehensive study, 2024.

[205] Stephen McAleese and Mark Keane. A comparative analysis of counterfactual explanation methods for text classifiers, 2024.

[206] Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. Zero-shot llm-guided counterfactual generation: A case study on nlp model evaluation, 2024.

[207] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

[208] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

[209] Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. Explaining the efficacy of counterfactually augmented data. *arXiv preprint arXiv:2010.02114*, 2020.

[210] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*, 2021.

[211] Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. Disco: Distilling counterfactuals with large language models. *arXiv preprint arXiv:2212.10534*, 2022.

[212] Rachneet Sachdeva, Martin Tutek, and Iryna Gurevych. Catfood: Counterfactual augmented training for improving out-of-domain performance and calibration. *arXiv preprint arXiv:2309.07822*, 2023.

[213] Tao Feng, Yicheng Li, Li Chenglin, Hao Chen, Fei Yu, and Yin Zhang. Teaching small language models reasoning through counterfactual distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5831–5842, 2024.

[214] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[215] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[216] make_moons — scikit-learn 1.6.1 documentation. `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html`. Accessed: 2025-05-16.

[217] Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2nd edition, 1998.

[218] Rajendra Bhatia. *Positive definite matrices*. Princeton university press, 2009.

[219] Daniel Gilo and Shaul Markovitch. A general search-based framework for generating textual counterfactual explanations, 2023.

[220] Stig Hellemans, Andres Algaba, Sam Verboven, and Vincent Ginis. Flexible counterfactual explanations with generative models, 2025.

[221] OpenAI (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[222] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

[223] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[224] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[225] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2018.

[226] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. *Proceedings of the 7th ACM conference on Recommender systems*, 2013.

[227] Yelp. Yelp open dataset, 2025. Accessed: 2025-05-15.

[228] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[229] Benoit Courty, Victor Schmidt, Laura Duboc, Lucile Demailly, and Quentin Lefevre. Codecarbon: Estimate the carbon footprint of your computing. `https://github.com/mlco2/codecarbon`, 2023. Version 2.8.5, Zenodo. DOI: 10.5281/zenodo.15239865.

[230] Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: a big data – ai integration perspective, 2019.

[231] Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H. Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. How to train data-efficient llms, 2024.

[232] Omar Y. Al-Jarrah, Paul D. Yoo, Sami Muhaidat, George K. Karagiannidis, and Kamal Taha. Efficient machine learning for big data: A review. *Big Data Research*, 2(3):87–93, 2015.

[233] Matt Bonakdarpour, Joshua Bon, and Matthew Stephens. Asymptotic normality of mle, 2019. Accessed: 2025-05-20.

[234] Qiang Yang. *Federated learning / Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, Han Yu.* Synthesis lectures on artificial intelligence and machine learning, #43. Morgan & Claypool, San Rafael, California, 2020.

[235] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[236] Megan Smith, Cecilia Munoz, and D. J. Patil. Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights, 2016.

[237] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021.

[238] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.

[239] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

[240] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 7494–7502, 2023.

[241] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems*, 34:26091–26102, 2021.

[242] Yuxin Shi, Han Yu, and Cyril Leung. A survey of fairness-aware federated learning. *arXiv preprint arXiv:2111.01872*, 2021.

[243] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

[244] Lin Wang, Zhichao Wang, and Xiaoying Tang. Fedeba+: Towards fair and effective federated learning via entropy-based model. *arXiv preprint arXiv:2301.12407*, 2023.

[245] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

[246] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 2016.

[247] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[248] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[249] Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. Fedfair: Training fair models in cross-silo federated learning. *arXiv preprint arXiv:2109.05662*, 2021.

[250] Borja Rodríguez-Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. *arXiv preprint arXiv:2109.08604*, 2021.

[251] Daniel Yue Zhang, Ziyi Kou, and Dong Wang. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1051–1060, 2020.

[252] Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Minimax demographic group fairness in federated learning. *arXiv preprint arXiv:2201.08304*, 2022.

[253] Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Provably fair federated learning via bounded group loss. *arXiv preprint arXiv:2203.10190*, 2022.

[254] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021.

[255] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

[256] S. Dutta, D. Wei, H. Yueksel, P. Y. Chen, S. Liu, and K. R. Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning (ICML)*, pages 2803–2813, 2020.

[257] Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. Fact: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, pages 5264–5274. PMLR, 2020.

[258] Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1):2527–2552, 2022.

[259] Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science*, 19(3):513–537, 2022.

[260] Hao Wang, Luxi He, Rui Gao, and Flavio Calmon. Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[261] Sivan Sabato and Elad Yom-Tov. Bounding the fairness and accuracy of classifiers from population statistics. In *International conference on machine learning*, pages 8316–8325. PMLR, 2020.

[262] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in classification. *arXiv preprint arXiv:1705.09055*, 2017.

[263] Praveen Venkatesh, Sanghamitra Dutta, Neil Mehta, and Pulkit Grover. Can information flows suggest targets for interventions in neural circuits? *Advances in Neural Information Processing Systems*, 34:3149–3162, 2021.

[264] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

[265] AmirEmad Ghassami, Sajad Khodadadian, and Negar Kiyavash. Fairness in supervised learning: An information theoretic approach. In *2018 IEEE international symposium on information theory (ISIT)*, pages 176–180, 2018.

[266] S. Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. An information-theoretic quantification of discrimination with exempt features. In *AAAI Conference on Artificial Intelligence*, 2020.

[267] Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. Fairness under feature exemptions: Counterfactual and observational measures. *IEEE Transactions on Information Theory*, 67(10):6675–6710, 2021.

[268] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using mutual information. In *2020 IEEE international symposium on information theory (ISIT)*, pages 2521–2526. IEEE, 2020.

[269] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Renyi fair inference. *arXiv preprint arXiv:1906.12005*, 2019.

[270] Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Fairness-aware neural reyni minimization for continuous features. *arXiv preprint arXiv:1911.04929*, 2019.

[271] Hao Wang, Hsiang Hsu, Mario Diaz, and Flavio P Calmon. To split or not to split: The impact of disparate treatment in classification. *IEEE Transactions on Information Theory*, 67(10):6733–6757, 2021.

[272] Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney. Causal feature selection for algorithmic fairness. In *Proceedings of the 2022 International Conference on Management of Data*, pages 276–285, 2022.

[273] Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. In *Advances in Neural Information Processing Systems*, volume 35, pages 38747–38760. Curran Associates, Inc., 2022.

[274] Peter Kairouz, Jiachun Liao, Chong Huang, Maunil Vyas, Monica Welfert, and Lalitha Sankar. Generating fair universal representations using adversarial models. *arXiv preprint arXiv:1910.00411*, 2019.

[275] David A Ehrlich, Andreas C Schneider, Michael Wibral, Viola Priesemann, and Abdullah Makkeh. Partial information decomposition reveals the structure of neural representations. *arXiv preprint arXiv:2209.10438*, 2022.

[276] Tycho MS Tax, Pedro AM Mediano, and Murray Shanahan. The partial information decomposition of generative neural network models. *Entropy*, 19(9):474, 2017.

[277] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36, 2023.

[278] Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *J. Mach. Learn. Res.*, 24:131–1, 2023.

[279] Salman Mohamadi, Gianfranco Doretto, and Donald A Adjeroh. More synergy, less redundancy: Exploiting joint mutual information for self-supervised learning. *arXiv preprint arXiv:2307.00651*, 2023.

[280] Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Michael Wibral, and Elad Schneidman. Estimating the unique information of continuous variables. *Advances in neural information processing systems*, 34:20295–20307, 2021.

[281] Praveen Venkatesh and Gabriel Schamberg. Partial information decomposition via deficiency for multivariate gaussians. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 2892–2897. IEEE, 2022.

[282] Mukund Prasad Sah and Amritpal Singh. Aggregation techniques in federated learning: Comprehensive survey, challenges and opportunities. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 1962–1967, 2022.

[283] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[284] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

[285] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[286] Artemy Kolchinsky. A novel approach to the partial information decomposition. *Entropy*, 24(3):403, 2022.

[287] R. G. James, C. J. Ellison, and J. P. Crutchfield. dit: a Python package for discrete information theory. *The Journal of Open Source Software*, 3(25):738, 2018.

[288] Tom Yan and Chicheng Zhang. Active fairness auditing. In *International Conference on Machine Learning*, pages 24929–24962. PMLR, 2022.

[289] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[290] Michael Kleinman, Alessandro Achille, Stefano Soatto, and Jonathan C Kao. Redundant information neural estimation. *Entropy*, 23(7):922, 2021.

[291] The White House. Blueprint for an ai bill of rights: Making automated systems work for the american people. `https://www.whitehouse.gov/ostp/ai-bill-of-rights/`, 2022. Accessed: [30 Jan, 2024].

[292] Jacy Anthis and Victor Veitch. Causal context connects counterfactual fairness to robust prediction and group fairness. *Advances in Neural Information Processing Systems*, 36, 2024.

[293] Anne L Washington. How to argue with an algorithm: Lessons from the compas-propublica debate. *Colo. Tech. LJ*, 17:131, 2018.

[294] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[295] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[296] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104:671, 2016.

[297] Carol Xuan Long, Hsiang Hsu, Wael Alghamdi, and Flavio Calmon. Individual arbitrariness and group fairness. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[298] Meiyu Zhong and Ravi Tandon. Intrinsic fairness-accuracy tradeoffs under equalized odds. *arXiv preprint arXiv:2405.07393*, 2024.

[299] Corinna Hertweck and Tim Räz. Gradual (in) compatibility of fairness criteria. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11926–11934, 2022.

[300] Brian Hsu, Rahul Mazumder, Preetam Nandy, and Kinjal Basu. Pushing the limits of fairness impossibility: Who's the fairest of them all? *Advances in Neural Information Processing Systems*, 35:32749–32761, 2022.

[301] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 3315–3323, 2016.

[302] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[303] Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Zakharchenko, Lucas Rosenblatt, and Julia Stoyanovich. The possibility of fairness: Revisiting the impossibility theorem in practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 400–422, 2023.

[304] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer, 2012.

[305] Hao Wang, Hsiang Hsu, Mario Diaz, and Flavio P. Calmon. The impact of split classifiers on group fairness. In *2021 IEEE International Symposium on Information Theory (ISIT)*, page 3179–3184. IEEE Press, 2021.

[306] Peter Kairouz, Jiachun Liao, Chong Huang, Maunil Vyas, Monica Welfert, and Lalitha Sankar. Generating fair universal representations using adversarial models. *IEEE Transactions on Information Forensics and Security*, 17:1970–1985, 2022.

[307] Barproda Halder, Faisal Hamman, Pasan Dissanayake, Qiuyi Zhang, Ilia Sucholutsky, and Sanghamitra Dutta. Quantifying spuriousness of biased datasets using partial information decomposition. *arXiv preprint arXiv:2407.00482*, 2024.

[308] Praveen Venkatesh, Keerthana Gurushankar, and Gabriel Schamberg. Capturing and interpreting unique information. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2631–2636, 2023.

[309] David Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pages 265–272, 1953.

[310] Pradeep Kr Banerjee, Eckehard Olbrich, Jürgen Jost, and Johannes Rauh. Unique informations and deficiencies. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 32–38, 2018.

[311] Masashi Sode Akihiko Fukuchi, Yoko Yabe. Fairtorch. `https://github.com/wbawakate/fairtorch`, 2021.

[312] Jianlong Zhou, Fang Chen, and Andreas Holzinger. Towards explainability for ai fairness. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 375–386. Springer, 2020.

[313] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust algorithmic recourse under model multiplicity with probabilistic guarantees. *IEEE Journal on Selected Areas in Information Theory*, pages 1–1, 2024.

[314] Praveen Venkatesh, Corbett Bennett, Sam Gale, Tamina K. Ramirez, Greggory Heller, Severine Durand, Shawn R Olsen, and Stefan Mihalas. Gaussian partial information decomposition: Bias correction and application to high-dimensional data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[315] Chaitanya Goswami, Amanda Merkley, and Pulkit Grover. Computing unique information for poisson and multinomial systems, 2023.

[316] Grigor Nalbandyan, Rima Shahbazyan, and Evelina Bakhturina. Score: Systematic consistency and robustness evaluation for large language models. *arXiv preprint arXiv:2503.00137*, 2025.

[317] Vidmantas Bentkus. On hoeffding's inequalities. *Annals of probability*, pages 1650–1673, 2004.

[318] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

[319] Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. *Advances in neural information processing systems*, 34:23928–23941, 2021.

[320] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[321] Pablo Zegers. Fisher information properties. *Entropy*, 17(7):4918–4939, 2015.

[322] Anand Avati. Bias-variance analysis: theory and practice, 2023.

[323] Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. Multifair: Multi-group fairness in machine learning. *arXiv preprint arXiv:2105.11069*, 2021.

[324] Ziv Goldfeld and Yury Polyanskiy. The information bottleneck problem and its applications in machine learning. *IEEE Journal on Selected Areas in Information Theory*, 1(1):19–38, 2020.

[325] Clément L Canonne. A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*, 2022.