Explaining Differences Between Model Pairs in Natural Language through Sample Learning

Advaith Malladi[1] **Rakesh R. Menon**[2] **Yuvraj Jain**[3] **Shashank Srivastava**[3]
[1]IIIT Hyderabad [2]Adobe Inc. [3]University of North Carolina at Chapel Hill
advaith.malladi@research.iiit.ac.in
rakeshrmenon1995@gmail.com
yjain@unc.edu, ssrivastava@cs.unc.edu

Abstract

With the growing adoption of machine learning models in critical domains, techniques for explaining differences between models have become essential for trust, debugging, and informed deployment. Previous approaches address this by identifying input transformations that cause divergent predictions (Shah et al., 2022) or by learning joint surrogate models to align and contrast behaviors (Haldar et al., 2023). These methods often require access to training data and do not produce natural language explanations. In this paper, we introduce SLED , *a framework that generates faithful natural language explanations of when and how two ML models converge or diverge in their predictions*. SLED first uses gradient-based optimization to synthesize input samples that highlight divergence and convergence patterns, and then leverages a large language model (LLM) to generate explanations grounded in these synthetic samples. Across both text-based (3 tasks, 7 models) and structured (10 tasks, 4 models) classification tasks, we show that SLED explanations are 18–24% more faithful than the strongest baselines. User studies also indicate that SLED explanations achieve a real-world simulatability of 63.5%. Importantly, SLED requires minimal access to training data and generalizes well to real-world samples, enabling transparent model comparison. [1]

1 Introduction

Machine learning models trained for the same task often exhibit surprising differences in behavior due to variations in training data, model architectures, or optimization techniques. For instance, consider two sentiment classifiers that were trained on slightly different review datasets. When presented with a complex movie review, Model A might label it as negative (recognizing subtle sarcasm), while Model B predicts positive (missing the sarcasm). Why do these models disagree? Understanding such differences is crucial for interpretability, model debugging, debiasing, and making informed deployment decisions. Communicating how two models diverge in plain language is particularly important, as textual explanations are accessible and user-friendly (Luo et al., 2024). However, generating these model divergence explanations is challenging, especially when we have limited access to the models' original training data. This raises the question: *Can we explain how two models behave differently with minimal access to their architecture or training data?*

In this work, we study the problem of *generating faithful natural language explanations that describe where a pair of models converge (agree) or diverge (disagree) in their predictions*. Throughout the paper, we refer to samples where the models agree in their predictions as convergent samples and where they disagree as divergent samples. Our goal here is to produce explanations that accurately reflect the models' differing decision logic in a human-interpretable way.

To interpret ML models, Ribeiro et al. (2016) and Ribeiro et al. (2018) generate local explanations using feature importance and high-precision rules, respectively, but they lack model-level insights. Menon et al. (2023) produce natural language explanations for model predictions but do not address divergence or convergence between models. Shah et al. (2022) focus on identifying input transformations that lead to divergent behavior, yet they do not provide textual explanations and are not applicable to text-based tasks.

We introduce SLED : Sample Learning to Explain Divergence Between Models, a framework that generates faithful natural language explanations describing where and how two ML models agree or disagree in their predictions. SLED op-

[1]Code and data for reproducing all experiments will be released on first publication.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28726–28748
November 4-9, 2025

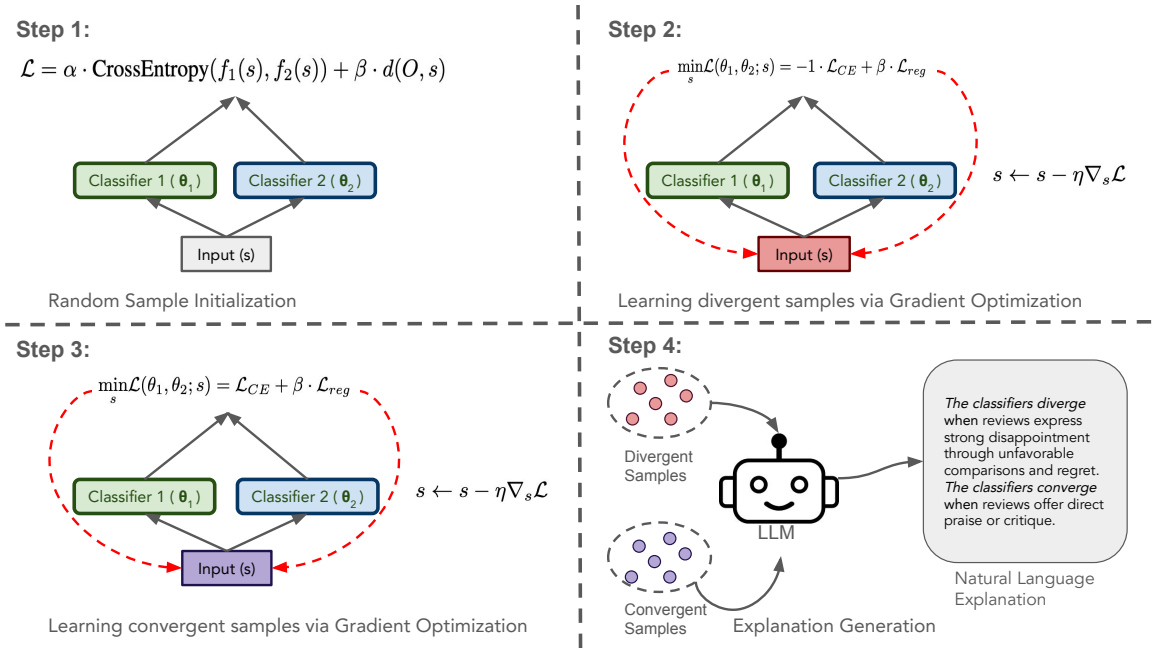

Figure 1: Overview of our SLED framework for explanations describing differences between pairs of models. The framework consists of four steps: (1) Initializing random input samples (2) Applying gradient updates on the random samples to learn synthetic divergent samples (3) Applying gradient updates to learn synthetic convergent samples (4) Prompting an **Explainer LLM** to generate a text-based explanation based on the synthetic samples.

erates in two stages: (1) learning synthetic convergent and divergent samples via gradient-based optimization with a distance regularization with the original data distribution, and (2) generating text-based explanations using these samples with an **Explainer LLM** (Figure 1).

We evaluate SLED on a variety of tasks and models. In experiments with both text classification tasks (spanning sentiment analysis and other NLP classification problems) and structured classification tasks (tabular benchmarks), SLED significantly outperforms prior approaches. For each input, we assess the quality of the explanation by using it to predict whether the models will agree or disagree (divergence prediction), with ground truth based on model outputs (§ 3.3). We measure faithfulness and simulatability (§ 3.3), comparing SLED to MaNtLE (Menon et al., 2023), LIME (Ribeiro et al., 2016), and Anchors (Ribeiro et al., 2018). On structured tasks, SLED is 24–48% more faithful, and on text tasks, it outperforms LLM-based explanations by 18%. Performance remains stable across different explainer LLMs and input sizes. We also show that distance regularization improves alignment with real data, significantly enhancing both faithfulness and simulatability (§ 5.3). Through user studies, we find that SLED explanations achieve a real-world simulatability of 63.5%. Human evaluations also reveal that our explanations are understandable, informative, and have a high perceived utility (§ 5.4).

Our contributions are: (1) we develop SLED , a novel framework that explains the divergence and convergence between machine learning models using faithful natural language explanations; (2) we design a synthetic sample learning method that generates synthetic samples revealing areas of divergence and convergence between two models while closely aligning with the original data distribution; (3) we demonstrate the effectiveness of SLED across multiple structured and text-based classification tasks with varied model configurations; (4) we show that SLED outperforms existing explanation frameworks (e.g., MaNtLE, LIME, Anchors) in generating more faithful model-level explanations; and (5) through user studies, we also demonstrate the effectiveness of SLED explanations in improving users' ability to simulate model predictions, understand model differences, and make informed decisions based on those explanations.

2 Related Work

Extensive research explains individual model predictions. *LIME* (Ribeiro et al., 2016) uses linear models to approximate local behavior and identify key features. *Anchors* (Ribeiro et al., 2018) extends this by creating precise, human-readable

rules ("anchors") that ensure consistency under perturbations. Both methods, though model-agnostic and effective for local interpretation, do not explain differences between models or offer natural language explanations.

Other works have proposed fine-tuning language models to generate explanations that align with classification outputs. *CAGE* (Rajani et al., 2019) and *WT5* (Narang et al., 2020) generate textual explanations that provide interpretable rationales. However, the explanations are limited to individual samples and do not offer insights at the model level. Menon et al. (2023) introduce *MaNtLE*, a model-agnostic explanation framework that uses large-scale pretraining on synthetic tasks to generate faithful natural language explanations. However, MaNtLE focuses on explaining single-model predictions and does not address model differences. To compare models, *ModelDiff* (Shah et al., 2022) identifies input transformations that induce divergent predictions. While effective for highlighting behavioral shifts, the outputs are not natural language explanations. Haldar et al. (2023) propose *Interpretable Differencing*, which learns a joint surrogate model to capture the behavioral gaps. Although suitable for structured data and useful for interpretation, the method assumes access to training data and is not applicable to text classification or settings with limited data.

3 SLED

In this section, we first define our problem setup, the various components of SLED , and the stages involved in generating explanations. Figure 1 illustrates our framework.

3.1 Problem Setup

We consider two machine learning classifiers, $f_1 : \mathcal{X} \to \mathcal{Y}$ and $f_2 : \mathcal{X} \to \mathcal{Y}$, parameterized by θ_1 and θ_2 respectively, trained for the same task $\mathcal{T}$. The input space $\mathcal{X}$ is defined by the nature of the task: if $\mathcal{T}$ is a structured prediction task, $\mathcal{X}$ contains structured inputs; if $\mathcal{T}$ is a text classification task, $\mathcal{X}$ contains natural language inputs. The label space $\mathcal{Y}$ consists of the labels for $\mathcal{T}$. Classifier f_1 is trained on dataset D_1, and classifier f_2 is trained on a separate dataset D_2, with both datasets corresponding to the same task $\mathcal{T}$. We define $O \subset \mathcal{X}$ as a small reference set comprising 10 randomly sampled inputs from the task domain. This set serves as an approximate representation of the input distribution of $\mathcal{T}$.

Algorithm 1 SLED

Require: Classifiers f_1, f_2; Data O; Task $\mathcal{T}$; Samples $\mathcal{N}$; Steps `max_steps`; Thresholds μ, σ; Distance d
Ensure: Explanation $\mathcal{E}$

if $\mathcal{T}$ is text **then**
  $O' \leftarrow \mathsf{Emb}_1(O)$; learn $\mathcal{A}$: $\mathsf{Emb}_1 \to \mathsf{Emb}_2$
else
  $\mathcal{A} \leftarrow$ Identity
end if
$Z_1, Z_2 \leftarrow \emptyset$
for $i = 1$ to $2\mathcal{N}$ **do**
  $\alpha \leftarrow (-1)^i$; $s \sim$ Random
  for $t = 1$ to `max_steps` **do**
    $\mathcal{L}_{\text{CE}} \leftarrow \alpha \cdot \text{CE}(f_1(s), f_2(\mathcal{A}(s)))$
    $\beta \leftarrow \begin{cases} = 10 & \text{if } d(O,s) < \mu + \sigma \\ = 0.05 & \text{otherwise} \end{cases}$
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}} + \beta \cdot d(O, s)$
    Update s via gradient step on $\mathcal{L}$
  end for
  if $\alpha = 1$ and $\arg\max f_1(s) = \arg\max f_2(s)$ and $d(O,s) < \mu + \sigma$ **then**
    $Z_2 \leftarrow Z_2 \cup \{s\}$
  else if $\alpha = -1$ and $\arg\max f_1(s) \neq \arg\max f_2(s)$ and $d(O,s) < \mu + \sigma$ **then**
    $Z_1 \leftarrow Z_1 \cup \{s\}$
  end if
end for
if $\mathcal{T}$ is text **then**
  Generate subcategories via LLM; synthesize corpus $\mathcal{C}$
  for $s \in Z_1 \cup Z_2$ **do**
    Replace s with nearest neighbor in $\mathcal{C}$
  end for
end if
$\mathcal{E} = LLM(Z_1, Z_2)$
return $\mathcal{E}$

3.2 Approach

Our framework consists of two primary stages: (1) synthetic sample learning, and (2) explanation generation. In the first stage, we generate two sets of synthetic inputs: a set of divergent (Z_1) and convergent (Z_2) samples. These sets are then passed to the explanation generation stage, which produces an explanation $\mathcal{E}$ that captures the divergence and convergence between the classifiers f_1 and f_2. By learning both divergent and convergent samples, we aim to generate a precise explanation $\mathcal{E}$ that differentiates these areas effectively. The algorithm for SLED can be found in Algorithm 1.

Synthetic Sample Learning for Structured Tasks

We use an iterative, gradient-based procedure to learn synthetic samples that capture model convergence or divergence. Starting from a randomly initialized sample s, we optimize a custom loss $\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{reg}}$, where:

A Cross-Entropy Loss, $\mathcal{L}_{\text{CE}}$ captures the divergence or convergence between the models. We define cross-entropy loss $\mathcal{L}_{\text{CE}} = \alpha \cdot CrossEntropy(f_1(s), f_2(s))$, the cross entropy from the output distribution of f_1 to f_2. By adjusting α, we control the behavior of the optimization:

- α = 1: minimizing cross entropy encourages convergence as optimization reduces the divergence between $f_1(s)$ and $f_2(s)$)
- α = -1: maximizing cross entropy encourages divergence as optimization increases the divergence between $f_1(s)$ and $f_2(s)$)

A Regularization Loss, $\mathcal{L}_{\text{reg}}$, keeps samples close to the original data distribution O, using the Mahalanobis distance $d(O, s)$: $\mathcal{L}_{\text{reg}} = \beta \cdot d(O, s)$ The scaling factor β is dynamically adjusted based on the current optimization state:

- If $d(O, s) > \mu + \sigma$, β is set to 10 to prioritize minimizing $\mathcal{L}_{\text{reg}}$.
- If $d(O, s) \leq \mu + \sigma$, β is set to 0.05 to prioritize minimizing $\mathcal{L}_{\text{CE}}$ because $d(O, s)$ is already within the desired range.

Here, μ and σ denote the mean and standard deviation of Mahalanobis distances computed across all samples in O.

Optimization terminates when the models converge ($\arg\max f_1(s) = \arg\max f_2(s)$) or diverge ($\arg\max f_1(s) \neq \arg\max f_2(s)$), and $d(O, s) < \mu + \sigma$. We generate sets of $\mathcal{N}$ convergent (Z_2) and divergent (Z_1) samples for explanation generation. The regularization loss ensures that all synthetic samples lie within a Mahalanobis distance of $\mu + \sigma$ from the original data distribution.

Synthetic Sample Learning for Text-based Tasks

Extending the same optimization procedure to text-based classifiers is non-trivial due to differences in tokenization schemes across models, which result in distinct embedding spaces. To bridge this gap, we learn an affine transformation $\mathcal{A}$ that maps embeddings from the space of f_1 to that of f_2. Specifically, we perform regression on aligned pairs of embeddings from both models and ensure that the learned transformation achieves an average R^2 score of at least 0.85 across all pairs, indicating a high-quality alignment. We then modify the cross-entropy loss as follows: $\mathcal{L}_{\text{CE}} = \text{CrossEntropy}(f_1(s), f_2(\mathcal{A}(s)))$. All other components remain unchanged. Using this objective, we learn a set of $\mathcal{N}$ convergent sentence embeddings Z_1 and $\mathcal{N}$ divergent sentence embeddings Z_2.

To map the sentence embeddings to text samples, we construct a synthetic corpus of task-relevant sentences and retrieve the nearest sentences to the embeddings in Z_1 and Z_2 using cosine similarity. The corpus is generated using a **Data Generator LLM**, following prior work showing that LLMs can produce high-quality synthetic data for classification tasks (Li et al., 2024, 2023). Given a classification task $\mathcal{T}$, we first prompt the LLM to enumerate representative subcategories that capture semantic or behavioral variations—e.g., in hate speech detection: *explicit hate speech*, *coded or euphemistic hate*, *neutral factual content*, and *sarcastic or ambiguous statements* (see Table 9 for prompt). These subcategories guide sentence generation. Using a set O of few-shot examples drawn from the task $\mathcal{T}$, we prompt the LLM to generate 50 new samples per subcategory (see Table 6 for prompt). The resulting sentences are aggregated into an unfiltered synthetic corpus. For each embedding in Z_1 and Z_2, we retrieve its nearest neighbor from the corpus to form the final decoded sets, which reflect the divergence and convergence patterns learned by SLED and serve as inputs for explanation generation. Since these synthetic samples are retrieved from a set of semantically valid, task-oriented examples generated by an LLM, this ensures that the samples are meaningful, human-interpretable, and aligned with the task.

Explanation Generation

Prior work has shown that LLMs can produce accurate and interpretable explanations from structured example sets by identifying and articulating underlying patterns (R Menon and Srivastava, 2024; Gat et al., 2023; Wang et al., 2024; Singh et al., 2022; Siegel et al., 2025). Building on this, we use an **Explainer LLM** to analyze the synthetic samples learned in the previous step and generate explanations that capture the behavioral differences between models.

In a zero-shot setting, we prompt the LLM to generate an explanation $\mathcal{E}$ that captures the areas of divergence and convergence between f_1 and f_2 based on the synthetic sample sets Z_1 and Z_2. The prompt is designed to elicit precise, faithful, pattern-level descriptions that generalize to unseen examples. The prompt for explanation generation

can be found in Table 7. Examples of explanations generated by SLED can be found in Table 1.

3.3 Evaluation Metrics

We evaluate the quality of our natural language explanations using two widely recognized metrics in explainable AI: *faithfulness* and *simulatability*. Faithfulness measures how well the explanation reflects the true behavior of the models on the examples used to generate the explanation (Jacovi and Goldberg, 2020). Simulatability assesses whether the explanation enables accurate prediction of model behavior on unseen samples (Hase and Bansal, 2020). Our explanations characterize the divergence and convergence between two classifiers. To evaluate them, we assess how well the explanation predicts whether the classifiers will converge or diverge on individual samples. Following recent work demonstrating the effectiveness of using LLMs as proxies for human evaluators in explainability tasks (Poché et al., 2025; Bona et al., 2024), we utilize a **Predictor LLM**, which is a proxy for a human user, to predict classifier convergence or divergence based solely on the explanation and the input sample. We measure faithfulness by evaluating how well the explanation captures true classifier behavior on the synthetic samples used to generate it. Specifically, we prompt the predictor LLM with each synthetic sample and the generated explanation, asking it to predict whether the classifiers will converge or diverge.

Faithfulness is computed as the fraction of predictions that match the actual convergence or divergence between the models. To measure simulatability, we evaluate the explanation's ability to generalize to a set of unseen samples. We ensure that this test set is balanced with respect to convergent and divergent samples, as simulatability scores should not reward degenerate explanation strategies (e.g., always predicting divergence). For each test sample, we prompt the predictor LLM with the explanation and the input, and record whether it correctly predicts model convergence or divergence. *Simulatability* is defined as the fraction of correct predictions on this balanced test set.

The evaluation of faithfulness was adapted to each method's explanation process using their best configurations. LIME and Anchors used real-world samples selected via sub-modular pick (SP), and faithfulness was measured on these. MaNtLE used its PF variant for better faithfulness, while SLED used synthetic samples from its optimization. This approach ensures each method is assessed fairly and accurately on the data it generated explanations from.

4 Experiments

In this section, we outline our experimental procedures to evaluate SLED explanations.

4.1 Tasks

Our evaluation covers two categories of tasks:

Text-based Classification tasks: We use 3 tasks: tweet hate-speech detection (Davidson et al., 2017), multi-class sentiment analysis (Parvez, 2023), and IMDB Movie Review polarity prediction (Maas et al., 2011)

4.2 Classifiers

SLED is designed to generate explanations for structured and text-based classification tasks. So, we include a diverse variety of ML models that can be utilized as classifiers. For our experiments with structured classification tasks, we use 4 models: linear regression model, logistic regression model, MLP classifier, and deep MLP classifier. For text-based classification tasks, we use the following 7 transformer (Vaswani et al., 2023) based models: bert-base-uncased and bert-large-uncased (Devlin et al., 2019), distilbert-base-uncased (Sanh et al., 2020), roberta-base (Liu et al., 2019), xlm-roberta-base (Conneau et al., 2020), llama-3.2-1B, and llama-3.2-1B Instruct (Grattafiori et al., 2024).

4.3 Baselines

Structured Classification Tasks: LIME approximates local model behavior by fitting a linear model to a classifier's predictions on perturbed inputs sampled near a given example. While originally designed for single-model explanations, we adapt LIME for model comparison by training a divergence classifier to predict whether two classifiers converge or diverge on a given input. This divergence classifier is trained on the full training set of the original models to ensure robustness. We then apply LIME to explain the divergence classifier's predictions, yielding interpretable explanations of model agreement and disagreement. LIME generates high-precision, human-readable rules. We apply Anchors to the same setting as above, to the divergence classifier, to provide rule-based interpretations of model convergence and divergence. We also evaluate both MaNtLE and a

Task	Explanation
Hate Speech Detection	Models diverge when the input expresses frustration, sarcasm, or ironic criticism of societal issues like hate speech or discrimination. Models converge when the input focuses on constructive dialogue, empathy, and actionable, respectful solutions.
Balance Scale	Models diverge when both distances are between 1 and 4, and at least one weight is low (1–3), without any side having maximum values. Models converge when one side has both distance and weight at 4 or 5, or when there is a clear extreme in distance or weight.

Table 1: Examples of SLED explanations for hate speech detection (text-based task) and balance scale (structured task). More examples can be found in Appendix D

	Faithfulness					Simulatability				
Task	SLED	**MaNtLE-PF**	**Mantle**	**LIME**	**Anchors**	SLED	**MaNtLE-PF**	**Mantle**	**LIME**	**Anchors**
Wine Quality	**0.97 ± 0.06**	0.50 ± 0.13	0.55 ± 0.13	0.35 ± 0.12	0.34 ± 0.20	**0.61 ± 0.21**	0.51 ± 0.17	0.53 ± 0.08	0.42 ± 0.12	0.32 ± 0.18
MAGIC Gamma	**0.95 ± 0.11**	0.50 ± 0.08	0.49 ± 0.14	0.48 ± 0.25	0.24 ± 0.18	**0.75 ± 0.14**	0.51 ± 0.06	0.48 ± 0.13	0.50 ± 0.20	0.24 ± 0.18
Rice	**0.99 ± 0.03**	0.52 ± 0.11	0.51 ± 0.14	0.53 ± 0.16	0.18 ± 0.16	**0.79 ± 0.16**	0.49 ± 0.13	0.49 ± 0.14	0.53 ± 0.12	0.26 ± 0.13
Bank Note	**0.69 ± 0.22**	0.53 ± 0.21	0.53 ± 0.10	0.46 ± 0.24	0.16 ± 0.18	**0.76 ± 0.20**	0.52 ± 0.18	0.52 ± 0.12	0.49 ± 0.23	0.18 ± 0.17
Adult	**0.84 ± 0.20**	0.55 ± 0.11	0.49 ± 0.27	0.49 ± 0.07	0.34 ± 0.14	**0.78 ± 0.11**	0.53 ± 0.12	0.48 ± 0.24	0.47 ± 0.09	0.37 ± 0.10
Bank Marketing	**0.68 ± 0.28**	0.60 ± 0.20	0.48 ± 0.15	0.50 ± 0.10	0.61 ± 0.42	0.65 ± 0.11	0.57 ± 0.21	0.48 ± 0.16	0.51 ± 0.06	**0.65 ± 0.43**
Car Evaluation	**0.53 ± 0.25**	0.47 ± 0.16	0.52 ± 0.09	0.50 ± 0.19	0.20 ± 0.28	**0.69 ± 0.16**	0.44 ± 0.19	0.50 ± 0.07	0.48 ± 0.16	0.20 ± 0.12
Tic Tac Toe	**0.53 ± 0.27**	0.49 ± 0.11	0.51 ± 0.11	0.48 ± 0.18	0.36 ± 0.19	**0.66 ± 0.06**	0.46 ± 0.09	0.50 ± 0.08	0.49 ± 0.14	0.30 ± 0.18
Nursery	0.55 ± 0.23	0.51 ± 0.11	0.50 ± 0.09	**0.59 ± 0.12**	0.20 ± 0.11	**0.73 ± 0.15**	0.52 ± 0.09	0.50 ± 0.12	0.50 ± 0.10	0.34 ± 0.16
Balance Scale	**0.85 ± 0.18**	0.50 ± 0.13	0.50 ± 0.14	0.48 ± 0.18	0.13 ± 0.14	**0.73 ± 0.16**	0.49 ± 0.09	0.51 ± 0.17	0.57 ± 0.10	0.20 ± 0.15
Overall	**0.76 ± 0.18**	0.52 ± 0.13	0.51 ± 0.14	0.49 ± 0.16	0.28 ± 0.20	**0.72 ± 0.13**	0.50 ± 0.13	0.50 ± 0.13	0.50 ± 0.13	0.31 ± 0.18

Table 2: Faithfulness and simulatability scores for explanations across 10 different tasks. Results are averaged over 130 runs per task. Bold numbers indicate the best scores for each metric on a particular dataset. Scores indicate mean ± standard deviation.Scores are aggregated across various model pair configurations.

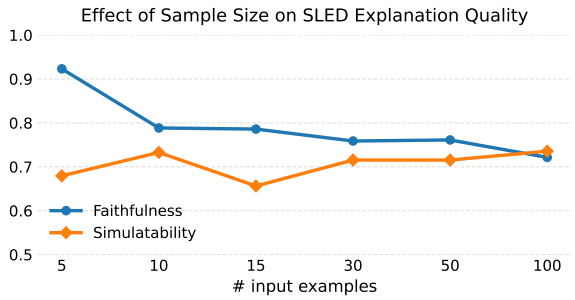

Figure 2: Effect of sample size on SLED explanation quality.

variant, MaNtLE-PF, which shows improved faithfulness. We generate explanations for both classifiers and concatenate them to create the final explanation.

Text-Based Classification Tasks: We use explanations generated by an explainer LLM with full access to the classifiers' training data as a baseline. We consider this a strong baseline given its unrestricted access to the data. LIME, Anchors, MaNtLE, and MaNtLE-PF were used as baselines only for structured classification tasks and were not used for text-based tasks, because of their limited applicability in text-based tasks.

5 Results and Analyses

In this section, we discuss and analyze results from experiments. For structured tasks, we report average performance across 130 runs, each employing different classifier configurations and training subsets. For text tasks, we average results over 100 runs. All classifiers are trained to convergence prior to explanation generation. We compare SLED against several baselines: MaNtLE-PF, MaNtLE, LIME (submodule pick), Anchors (submodule pick), and for text tasks, an LLM with full access to training data.

5.1 Faithfulness

SLED explanations are significantly more faithful to divergent and convergent model behavior than all baselines. On structured tasks (Table 2), SLED explanations achieve significantly higher faithfulness scores: 24% more than MaNtLE-PF, 25% more than MaNtLE, 27% more than LIME and 48% more than Anchors. In text classification tasks (Table 3), SLED explanations are 18% more faithful than the LLM baseline. SLED explanations demonstrate consistent faithfulness across all tasks. These improvements are statistically significant (paired t-test, $p < 0.01$), confirming that SLED , through its use of both divergent and conver-

Task	Faithfulness		Simulatability	
	SLED	LLM Explanations	SLED	LLM Explanations
Hate Speech Detection	**0.78 ± 0.11**	0.66 ± 0.10	**0.70 ± 0.12**	0.63 ± 0.08
IMDB Movie Reviews	**0.81 ± 0.10**	0.54 ± 0.06	**0.67 ± 0.11**	0.54 ± 0.09
Sentiment Analysis	**0.74 ± 0.12**	0.47 ± 0.08	**0.67 ± 0.12**	0.38 ± 0.12
Overall	**0.77 ± 0.11**	0.59 ± 0.13	**0.67 ± 0.16**	0.57 ± 0.16

Table 3: Faithfulness and simulatability scores across 3 tasks, comparing SLED and LLM-generated explanations. Results are averaged over 100 runs. Bold values indicate the best performance for each task. Scores indicate mean ± standard deviation.Scores are aggregated across various model pair configurations.

Explainer LLM	Faithfulness	Simulatability
GPT-4o mini	0.74 ± 0.14	0.74 ± 0.14
GPT-4.1 mini	0.77 ± 0.27	0.71 ± 0.15
GPT-4o	0.76 ± 0.18	0.72 ± 0.13
GPT-4.1	0.85 ± 0.20	0.74 ± 0.13

Table 4: Faithfulness and Simulatability of SLED Explanations across various explainer LLMs. GPT-4o was used for all experiments in Tables 2 and 3. Scores indicate mean ± standard deviation.

gent synthetic samples,‘ more accurately captures the true behavior of the models under comparison.

5.2 Simulatability

SLED explanations reliably capture and generalize the convergence and divergence behavior of classifiers on real-world samples. Tables 2 and 3 show that SLED explanations significantly outperform all other approaches in terms of simulatability (paired t-test, $p < 0.01$), indicating robust generalization to unseen real-world data. In particular, we note that SLED explanations achieve simulatability scores that are 22% higher than MaNtLE-PF, MaNtLE, and LIME, and 41% higher than Anchors. Similarly, in text-based tasks, the performance from SLED exceeds the LLM baseline performance by 10% in simulatability. These results demonstrate that the patterns captured by SLED are not only faithful on synthetic samples but also translate into more accurate predictive alignment on new, unseen data. Next, we look at a series of ablation studies that provide insights into SLED 's performance.

5.3 Ablations

How does the choice of the Explainer LLM affect the quality of the explanations?

We evaluate the impact of varying the explainer LLM used to generate SLED explanations, while keeping all other components constant. All standard experiments reported in Tables 2 and 3 were conducted using GPT-4o.

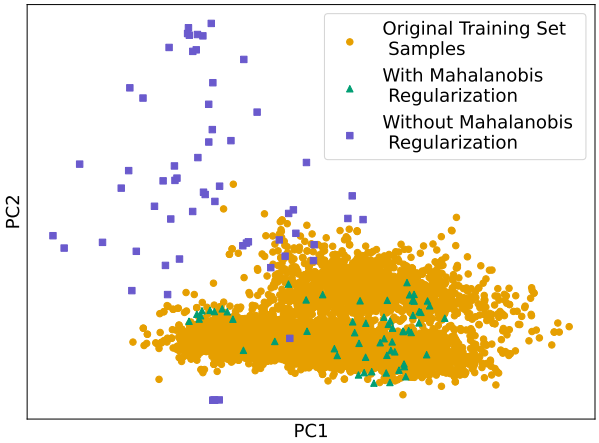

Figure 3: PCA projections of training set examples and SLED learned samples, with and without regularization.

In Table 4, we observe that faithfulness scores range from 0.74 (GPT-4o mini) to 0.85 (GPT-4.1). Using a smaller LLM than GPT-4o does not significantly affect the faithfulness score (paired t-test; p-value > 0.01). However, employing a more capable LLM like GPT-4.1 leads to higher faithfulness. Simulatability scores also remain within a narrow range (0.71–0.74) across all explainer models, suggesting that the generated explanations generalize equally well to unseen test instances. The consistency of both metrics indicates that SLED is robust to the choice of explainer LLM. Even smaller models like GPT-4o mini are sufficient for generating high-quality explanations, making SLED practical for low-resource settings.

How does the number of samples used for generating the explanation affect its quality?

To examine how the number of input examples influences explanation quality, we vary the number

Task Name	SLED Explanations		Without Regularization Loss	
	Faithfulness	Simulatability	Faithfulness	Simulatability
Wine Quality Dataset	**0.97 ± 0.05**	**0.61 ± 0.21**	0.85 ± 0.10	0.53 ± 0.16
MAGIC Gamma Telescope	**0.95 ± 0.11**	**0.75 ± 0.14**	0.71 ± 0.18	0.50 ± 0.05
Rice	**0.99 ± 0.03**	**0.79 ± 0.16**	0.92 ± 0.07	0.53 ± 0.09
Bank Note Auth	**0.69 ± 0.22**	**0.75 ± 0.21**	0.60 ± 0.23	0.63 ± 0.20
Adult	**0.84 ± 0.20**	**0.78 ± 0.11**	0.51 ± 0.21	0.53 ± 0.07
Bank Marketing	**0.68 ± 0.28**	**0.65 ± 0.11**	0.52 ± 0.20	0.51 ± 0.03
Car Evaluation	**0.54 ± 0.25**	**0.70 ± 0.16**	0.39 ± 0.17	0.50 ± 0.07
Tic Tac Toe	**0.53 ± 0.28**	**0.66 ± 0.06**	0.45 ± 0.23	0.49 ± 0.08
Nursery	**0.55 ± 0.24**	**0.73 ± 0.15**	0.52 ± 0.19	0.50 ± 0.11
Balance Scale	**0.85 ± 0.18**	**0.73 ± 0.16**	0.63 ± 0.20	0.50 ± 0.09

Table 5: Faithfulness and simulatability of SLED explanations with and without regularization loss. Scores indicate mean ± standard deviation.

of synthetic convergent and divergent samples provided to the explanation model and measure the resulting faithfulness and simulatability scores. As shown in Figure 2, we observe a clear trend: increasing the number of input examples beyond 50 does not improve explanation faithfulness. Interestingly, very high faithfulness scores are achieved even with as few as 5 examples. This may be because the LLM tends to explicitly describe each sample in the explanation rather than abstracting general patterns of convergence and divergence. Providing more than 50 examples appears to introduce noise, leading to a decline in faithfulness. Notably, explanations generated with as few as 30 examples yield competitive faithfulness and near-maximal simulatability.

How does the regularization loss impact the quality of the synthetic samples and the explanation?

To assess the impact of the regularization loss $\mathcal{L}_{\text{reg}}$ on generating high-quality synthetic samples, we generate SLED explanations for all structured classification tasks *without* applying $\mathcal{L}_{\text{reg}}$. As shown in Table 5, explanation quality drops significantly when the regularization component is removed. Specifically, we observe a 14% decrease in faithfulness and a 19% drop in simulatability. Figure 3 further demonstrates that omitting the regularization loss causes synthetic samples to fall outside the original training data distribution. In such cases, SLED tends to produce edge-case examples. In contrast, applying regularization yields samples that better reflect the training distribution, resulting in more coherent and faithful explanations.

5.4 Human Evaluation

To evaluate the quality of SLED explanations, we assess their understandability, informativeness, and perceived utility. We also examine whether humans can use SLED explanations to accurately predict model divergence or convergence.

We define **understandability** as how clear and comprehensible the explanations are in the given context. **Informativeness** measures whether the explanations reveal meaningful patterns of convergence and divergence between the models, rather than merely describing the provided samples. **Perceived utility** reflects participants' confidence in classifying new samples as divergent or convergent based on the explanation.

We recruited 10 participants, all pursuing at least an undergraduate degree in Computer Science, to rate SLED explanations on a 1–5 Likert scale for understandability, informativeness, and perceived utility. Additionally, participants were shown SLED explanations and asked to predict whether the models would converge or diverge on individual samples. We find that SLED explanations achieve a simulatability score of 63.5% with real human users. On average, participants rated the explanations highly: understandability at 3.86/5, informativeness at 4.61/5, and perceived utility at 3.87/5.

6 Conclusion

We present SLED , a framework for generating faithful natural language explanations of where two machine learning models converge and diverge, with minimal access to training data. SLED combines gradient-based generation of synthetic con-

vergent/divergent samples with LLM-based explanation generation to produce high-quality, faithful explanations. On evaluations across 13 datasets and 11 model configurations, SLED consistently outperforms baselines such as MaNtLE, LIME, Anchors, and explainer LLMs with access to the training set, achieving 18–24% improvements in faithfulness and 10–22% gains in simulatability. Through ablation studies, we demonstrate that: (1) a regularization loss encourages realistic synthetic samples and enhances explanation quality; (2) SLED is sample-efficient, requiring as few as 10 examples; and (3) explanation quality is robust to the choice of LLM, including smaller models. User studies also reveal that SLED explanations achieve a real-world simulatability score of 63.5%. Participants also found SLED explanations understandable, informative, and useful in classifying new samples

Looking ahead, future work can include extending SLED to multi-model tasks, and interactive settings where users can pose follow-up "why" questions on-the-fly. Integrating stronger safeguards against potential misuse, such as revealing sensitive failure modes, is another important direction.

Limitations

SLED is only applicable to ML models that are differentiable because of our reliance on gradient-based optimization to learn synthetic samples. Our framework also heavily relies on the capabilities of LLMs. The diversity and coverage of synthetic inputs for the text-based tasks depend on the LLM's ability to generate realistic and varied task-relevant samples. We evaluate SLED only on classification tasks; applicability to generative tasks remains unexplored. If the synthetic samples generated by SLED are overly similar to the training data, the resulting explanations may fail to capture model behavior on truly out-of-distribution inputs. In addition, our human evaluation was limited in demographic diversity, and a broader participant base would have been more beneficial.

Ethics Statement

This work uses only publicly available datasets and synthetically generated data, ensuring that no personally identifiable or sensitive information is involved. Human participants in evaluation studies were compensated fairly. While our framework focuses on explaining model behavior, it inherits limitations and potential biases from the underlying pre-trained models, which are beyond the scope of this study.

Acknowledgements

The authors would like to thank the anonymous reviewers for their suggestions and feedback on the work. This work was supported in part by NSF grant DRL2112635. The views contained in this article are those of the authors and not of the funding agency.

References

Francesco Bombassei De Bona, Gabriele Dominici, Tim Miller, Marc Langheinrich, and Martin Gjoreski. 2024. Evaluating explanations through llms: Beyond traditional user studies. *Preprint*, arXiv:2410.17781.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2023. Faithful explanations of black-box nlp models using llm-generated counterfactuals. *arXiv preprint arXiv:2310.00603*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Swagatam Haldar, Diptikalyan Saha, Dennis Wei, Rahul Nair, and Elizabeth M. Daly. 2023. Interpretable differencing of machine learning models. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in*

Artificial Intelligence, volume 216 of *Proceedings of Machine Learning Research*, pages 788–797. PMLR.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Yinheng Li, Rogerio Bonatti, Sara Abdali, Justin Wagle, and Kazuhito Koishida. 2024. Data generation using large language models for text classification: An empirical case study. *arXiv preprint arXiv:2407.12813*.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. 2024. Local interpretations for explainable natural language processing: A survey. *ACM Computing Surveys*, 56(9):1–36.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Rakesh Menon, Kerem Zaman, and Shashank Srivastava. 2023. MaNtLE: Model-agnostic natural language explainer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13493–13511, Singapore. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *Preprint*, arXiv:2004.14546.

Shahriar Parvez. 2023. multiclass-sentiment-analysis-dataset. `https://huggingface.co/datasets/Sp1786/multiclass-sentiment-analysis-dataset`.

Antonin Poché, Alon Jacovi, Agustin Martin Picard, Victor Boutin, and Fanny Jourdan. 2025. Consim: Measuring concept-based explanations' effectiveness with automated simulatability. *Preprint*, arXiv:2501.05855.

Rakesh R Menon and Shashank Srivastava. 2024. DISCERN: Decoding systematic errors in natural language for text classifiers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19565–19583, Miami, Florida, USA. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Harshay Shah, Sung Min Park, Andrew Ilyas, and Aleksander Madry. 2022. Modeldiff: A framework for comparing learning algorithms. *arXiv preprint arXiv:2211.12491*.

Noah Y. Siegel, Nicolas Heess, Maria Perez-Ortiz, and Oana-Maria Camburu. 2025. Faithfulness of llm self-explanations for commonsense tasks: Larger is better, and instruction-tuning allows trade-offs but not pareto dominance. *arXiv preprint arXiv:2503.13445*.

Chandan Singh, John X. Morris, Jyoti Aneja, Alexander M. Rush, and Jianfeng Gao. 2022. Explaining patterns in data with language models via interpretable autoprompting. *arXiv preprint arXiv:2210.01848*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Ruochen Wang, Si Si, Felix Yu, Dorothea Wiesmann, Cho-Jui Hsieh, and Inderjit Dhillon. 2024. Large language models are interpretable learners. *arXiv preprint arXiv:2406.17224*.

Prompt Template for Example Generation

Given a subcategory, your task is to generate 30 new sentences that exemplify it. Ensure that the generated sentences are human-like and closely match the style and register of the following examples:
{few_shot_examples}
The generated samples should strictly adhere to the given subcategory and follow the same style and format as the examples.
Subcategory: {subcategory}
Sentences:

Table 6: Prompt used to generate sentence examples for a given subcategory, based on few-shot examples.

Appendix

In the appendix, we add the prompts used across all LLMs, hyperparameter details, libraries, and examples of explanations generated using SLED .

A Prompts

The prompts for: generating synthetic samples can be found in Table 6, generating explanations can be found in Table 7,and generating subcategories can be found in Table 9

B Hyperparameter Details

Here, we outline the hyperparameters that have been used across various stages of SLED . Table 8 includes details about the sample learning algorithm, training classifiers, LLMs used, and hardware requirements.

C Libraries

We implement all classifiers using PyTorch. The sample learning stage of SLED was also implemented using PyTorch. We use the Huggingface library for all pre-trained models. Classifiers were trained with full precision until their performance could no longer be improved. We use the OpenAI API to make calls to GPT-4o mini, GPT-4o, GPT-4.1 mini, and GPT-4.1.

D Examples of SLED Explanations

Wine Quality

The models diverge when the combination of volatile acidity is at least 0.08 and at most 0.554, citric acid is at most 0.654, and pH is at least 2.769 and at most 3.7778, with at least one of the following: total sulfur dioxide is at most 294.18 and free sulfur dioxide is at most 85.96, or density is at least 0.9892 and at most 1.0052 with residual sugar at least 1.05 and at most 26.81, and sulphates is at least 0.22 and at most 1.147, while alcohol is at least 8.0 and at most 9.3; these conditions interact such that moderate to high acidity, variable sugar, and a wide pH range combine with either low to moderate sulphur dioxide or higher density and sulphates, creating ambiguous chemical profiles. The models converge when volatile acidity is at least 0.0845 and at most 0.5809, citric acid is at least 0.00498 and at most 0.654, pH is at least 2.775 and at most 3.5107, and the interaction of total sulfur dioxide at least 7.73 and at most 294.47 with free sulfur dioxide at least 1.29 and at most 85.96, density at least 0.9900 and at most 1.0052, residual sugar at least 1.05 and at most 26.81, sulphates at least 0.248 and at most 0.846, and alcohol at least 8.0069 and at most 11.2154, results in chemical profiles where the relationships among acidity, sulphur compounds, and sugar content are more consistent and less ambiguous, leading to model agreement.

MAGIC Gamma Telescope

The models diverge when the combination of fDist less than or equal to 250 and either fLength less than or equal to 35 or fWidth less than or equal to 12 is present, especially when these are paired with fSize less than or equal to 2.7 and at least one of the following: fConc greater than or equal to 0.43, fConc1 greater than or equal to 0.29, or fAlpha greater than or equal to 66. In these cases, the interaction of compact spatial features, moderate to high concentration values, and high orientation angles leads to disagreement. The models converge when fDist is greater than 250 regardless of other features, or when fDist is less than or equal to 250 but both fLength and fWidth exceed 35 and 12 respectively, or when fSize is greater than 2.7 and fConc and fConc1 are both less than 0.43 and 0.29, with fAlpha below 66, resulting in agreement due to the presence of more distributed spatial characteristics, lower concentration, and moderate orientation.

Rice

The models diverge when the combination of a Major Axis Length between 145 and 232, a Minor Axis Length below 90, and an Eccentricity above

Prompt Template for Explanation Generation

You are given two sets of samples:
Divergent Samples:
{divergent_list}
Convergent Samples:
{convergent_list}
Divergent samples are those where two machine learning models produce different predictions.
Convergent samples are those where the models produce the same predictions.
Your task is to generate a natural language explanation that identifies the general patterns or characteristics that distinguish divergent samples from convergent ones.
The explanation must:

- Clearly describe the shared properties or features that cause the models to diverge.
- Clearly describe the shared properties or features that lead the models to converge.
- Ensure that the identified divergence patterns apply to **all** divergent samples and **none** of the convergent ones.
- Ensure that the convergence patterns apply to **all** convergent samples and **none** of the divergent ones.
- Be generalizable — it should apply to unseen samples with similar characteristics.

Format your explanation as follows:
"The models diverge when ... The models converge when ..."
Do not refer to or quote the samples directly. Focus only on the abstract, generalizable patterns behind the divergence and convergence. Generate a single short paragraph with a detailed, precise, and human-understandable explanation.

Table 7: Prompt used to generate a generalizable explanation for model divergence and convergence based on synthetic samples generated by SLED .

0.93 occurs together, particularly when Area is between 7,900 and 16,500 and Extent is below 0.7, indicating elongated shapes with narrow widths and high eccentricity, often accompanied by relatively low Extent values. The models converge when either the Minor Axis Length is at least 90 regardless of Eccentricity, or when Eccentricity is below 0.93 even if the Minor Axis Length is less than 90, and these patterns are found across a broad range of Area and Extent values, reflecting more balanced or less elongated shapes where the relationship between axis lengths and eccentricity does not reach the critical thresholds that trigger disagreement.

Bank Note

The models diverge when variance is greater than or equal to 0, regardless of the values of skewness, curtosis, and entropy, or when variance is less than 0 but entropy is greater than or equal to 0, indicating that either a non-negative variance or the combination of negative variance with non-negative entropy leads to disagreement. The models converge when variance is less than 0 and entropy is less than 0, showing that only the joint presence of negative variance and negative entropy produces consistent agreement, regardless of the values of skewness and curtosis.

Bank Marketing

The models diverge when the individual is married, has an education level of 'tertiary', and is contacted in the months of May or June, with age between 18 and 34.5, duration less than or equal to 191.8, and previous contacts less than or equal to 2.2, regardless of balance, job, or housing status, and with 'poutcome' being either 'success' or 'other'. The models converge when these conditions are not simultaneously met, specifically when education is

Component	Details
Sample Learning	
If $d > u + \sigma$	$\beta = 10$
If $d < u + \sigma$	$\beta = 0.05$
Convergent samples	$\alpha = 1$
Divergent samples	$\alpha = -1$
# Divergent samples	$\mathcal{N} = 30$
# Convergent samples	$\mathcal{N} = 30$
Optimizer	Adam
Adam Beta1	0.9
Adam Beta2	0.999
Classifier Settings	
Learning Rate	1e-5
Batch Size	10
LR Scheduler	Linear
Warmup Steps	500
LLMs Used	
Explainer LLM	GPT-4o
Predictor LLM	GPT-4o
Data Generator LLM	GPT-4.1
Hardware	2× A6000 GPUs (48 GB VRAM each)

Table 8: Training settings, LLM configuration, and hardware requirements

not 'tertiary' or marital status is not 'married', or when age is above 34.5, or when duration exceeds 191.8, or when previous contacts are greater than 2.2, or when the month is not May or June, or when 'poutcome' is 'failure' or 'other' in the absence of the other criteria, reflecting a holistic interaction where the combination of marital status, education, age, timing, and engagement history determines model agreement.

Tic Tac Toe

The models diverge when the board contains at least five squares marked with 'x', at least three squares marked with 'b', and the 'middle-middle-square' is always 'o', with no row, column, or diagonal fully occupied by a single symbol; this configuration creates ambiguous board states where the distribution of 'x', 'o', and 'b' prevents a clear win or block, leading to different model interpretations. The models converge when the board features a mix of 'x', 'o', and 'b' such that the 'middle-middle-square' is not consistently 'o', and the arrangement allows for either a clear line of three identical symbols or an unambiguous path to victory or block, resulting in both models producing the same prediction due to the determinacy of the board state.

Nursery

The models diverge when the combination of 'housing' is 'convenient' or 'less conv', 'finance' is 'convenient', 'social' is 'nonprob', and 'health' is either

Prompt Template for Subcategory Generation

Given a classification task. Generate a diverse and representative set of fine-grained subcategories that capture distinct behavioral or semantic variations within the task domain. The subcategories should capture meaningful differences and variations in context, expression, and intent. The subcategories should capture all polarities involved in the task. Make sure the list of subcategories you generate is exhaustive. The list you generate should contain at least 10 subcategories.
For example:
Task: Hate Speech Detection
Subcategories: {Examples of subcategories in hate speech detection}
Now:
Task: Task T
Subcategories:

Table 9: Prompt used to generate fine-grained, polarity-aware subcategories for classification tasks.

'recommended' or 'priority', regardless of the values of 'parents', 'has nurs', 'form', or 'children', or when 'housing' is 'convenient', 'finance' is 'inconv', 'social' is 'slightly prob', and 'health' is 'priority'. The models converge when any of these feature interactions are not present, such as when 'housing' is 'critical', or when 'finance' is 'inconv', or when 'social' is not 'nonprob', or when 'health' is 'not recom', or when the combination of 'housing', 'finance', 'social', and 'health' does not match the specific patterns described for divergence.

Balance Scale

The models diverge when both the right-distance and left-distance are within the range of 1 to 4 and at least one of the weights (right-weight or left-weight) is 1, 2, or 3, with no side having both distance and weight simultaneously at their maximum values of 5; this pattern often involves combinations where neither side dominates in both distance and weight, and low or moderate values are distributed across features. The models converge when at least one side—either right or left—has both distance and weight at 4 or 5, or when the interaction of distances and weights across both

sides includes at least one feature at its maximum or minimum (1 or 5), creating a clear dominance or balance that leads to consistent model predictions.

IMDB Movie Reviews

The models diverge when the input contains strong emotional reactions to films, offering either extreme praise or harsh criticism. These inputs emphasize a clear sentiment and often highlight a stark contrast between outstanding acting and a weak plot. The models converge when the input takes a more balanced tone, acknowledging both strengths and weaknesses in the film. These inputs provide nuanced critiques, recognizing redeeming qualities alongside flaws without intense emotional polarization.

Sentiment Analysis

The models diverge when the input expresses personal experiences and emotional reactions, mixing positive and negative sentiments about situations like receiving packages, attending events, or using technology. These inputs highlight subjective viewpoints and individual feelings such as frustration, disappointment, or satisfaction. The models converge when the input presents neutral, factual statements or descriptions of mundane activities without emotional depth. These inputs lack personal reflection and focus on objective reporting rather than subjective experiences.

Hate Speech Detection

The models diverge when the input expresses frustration, sarcasm, or criticism toward societal issues such as hate speech, discrimination, and cultural behaviors. These inputs often use irony or humor to highlight hypocrisy or absurdity in social attitudes. The models converge when the input centers on constructive discussions, advocacy, and respectful dialogue. These inputs emphasize unity, empathy, and actionable solutions to address hate and discrimination without using sarcasm or negativity.

E User Study

The following questions were presented to the participants:

- **Simutalability**: Based on the provided explanation, classify the sample as convergent or divergent.
- **Understandability**: On a scale of 1-5, how understandable are these explanations?
- **Informativeness**: On a scale of 1-5, how useful are these explanations in capturing divergence/convergence patterns across models?
- **Percieved Utility**: On a scale of 1-5, how confident are you that you can classify a new samples as divergent or convergent based on the provided explanation.

F Synthetic Samples and Explanations

F.1 Wine Quality

Faithfulness: 0.84

Explanation: The classifiers diverge when encountering samples that exhibit a distinctive profile characterized by a fixed acidity ranging from 5.13 to 11.40, with a notable presence of citric acid between 0.21 and 0.77, which contributes to a balanced flavor profile. These samples typically have a volatile acidity between 0.1 and 0.4, suggesting a moderate level of acidity that complements the overall taste. Residual sugar levels are consistently elevated, spanning from 13.31 to 25.44, indicating a sweeter profile that harmonizes with the acidity. Chlorides remain low, generally under 0.04, while free sulfur dioxide levels vary from 3.3 to 82.79, reflecting a careful preservation approach. The total sulfur dioxide content ranges from 8.17 to 283.33, supporting the wine's stability and longevity. Density values hover around 0.99 to 1.01, and pH levels are maintained between 2.74 and 3.31, reinforcing the wine's freshness and crispness. Alcohol content is moderate, ranging from 8.01 to 12.20, which aligns well with the overall flavor profile, creating a well-rounded and enjoyable tasting experience. This combination of features creates a unique and appealing character that distinctly defines the positive samples, making them easily identifiable and classifiable. The classifiers converge when samples do not fit this distinctive profile.

Divergent Samples:

- {fixed_acidity: 6.11, volatile_acidity: 0.16, citric_acid: 0.21, residual_sugar: 21.2, chlorides: 0.01, free_sulfur_dioxide: 48.52, total_sulfur_dioxide: 260.76, density: 1.0, pH: 2.91, sulphates: 0.26, alcohol: 8.03}
- {fixed_acidity: 9.18, volatile_acidity: 0.27, citric_acid: 0.48, residual_sugar: 21.4, chlorides: 0.03, free_sulfur_dioxide: 23.18, total_sulfur_dioxide: 98.44, density: 1.01, pH: 2.94, sulphates: 0.49, alcohol: 8.28}

- {fixed_acidity: 8.52, volatile_acidity: 0.1, citric_acid: 0.77, residual_sugar: 16.9, chlorides: 0.01, free_sulfur_dioxide: 37.58, total_sulfur_dioxide: 146.18, density: 1.0, pH: 2.74, sulphates: 0.35, alcohol: 8.01}
- {fixed_acidity: 9.06, volatile_acidity: 0.22, citric_acid: 0.56, residual_sugar: 14.23, chlorides: 0.02, free_sulfur_dioxide: 17.7, total_sulfur_dioxide: 135.77, density: 1.0, pH: 2.89, sulphates: 0.29, alcohol: 8.01}
- {fixed_acidity: 5.82, volatile_acidity: 0.11, citric_acid: 0.41, residual_sugar: 14.42, chlorides: 0.01, free_sulfur_dioxide: 61.48, total_sulfur_dioxide: 251.64, density: 0.99, pH: 2.85, sulphates: 0.26, alcohol: 8.13}
- {fixed_acidity: 11.07, volatile_acidity: 0.29, citric_acid: 0.74, residual_sugar: 16.97, chlorides: 0.04, free_sulfur_dioxide: 12.23, total_sulfur_dioxide: 121.01, density: 1.0, pH: 3.05, sulphates: 0.55, alcohol: 9.28}
- {fixed_acidity: 8.93, volatile_acidity: 0.15, citric_acid: 0.63, residual_sugar: 15.92, chlorides: 0.02, free_sulfur_dioxide: 16.84, total_sulfur_dioxide: 109.73, density: 1.0, pH: 2.83, sulphates: 0.31, alcohol: 8.37}
- {fixed_acidity: 8.58, volatile_acidity: 0.1, citric_acid: 0.52, residual_sugar: 22.83, chlorides: 0.01, free_sulfur_dioxide: 47.08, total_sulfur_dioxide: 270.74, density: 1.0, pH: 2.83, sulphates: 0.29, alcohol: 8.03}
- {fixed_acidity: 9.51, volatile_acidity: 0.4, citric_acid: 0.58, residual_sugar: 13.71, chlorides: 0.02, free_sulfur_dioxide: 3.3, total_sulfur_dioxide: 163.11, density: 1.0, pH: 3.0, sulphates: 0.3, alcohol: 8.04}
- {fixed_acidity: 9.62, volatile_acidity: 0.22, citric_acid: 0.61, residual_sugar: 17.62, chlorides: 0.01, free_sulfur_dioxide: 34.98, total_sulfur_dioxide: 216.92, density: 1.0, pH: 2.81, sulphates: 0.25, alcohol: 8.01}

Convergent Samples:

- {fixed_acidity: 10.71, volatile_acidity: 0.53, citric_acid: 0.56, residual_sugar: 10.31, chlorides: 0.1, free_sulfur_dioxide: 5.03, total_sulfur_dioxide: 61.99, density: 1.0, pH: 2.96, sulphates: 0.64, alcohol: 8.07}
- {fixed_acidity: 11.33, volatile_acidity: 0.13, citric_acid: 0.51, residual_sugar: 1.51, chlorides: 0.02, free_sulfur_dioxide: 1.86, total_sulfur_dioxide: 7.3, density: 0.99, pH: 2.84, sulphates: 0.32, alcohol: 9.95}
- {fixed_acidity: 10.04, volatile_acidity: 0.47, citric_acid: 0.36, residual_sugar: 9.34, chlorides: 0.14, free_sulfur_dioxide: 5.9, total_sulfur_dioxide: 12.51, density: 1.0, pH: 2.92, sulphates: 0.64, alcohol: 8.03}
- {fixed_acidity: 6.69, volatile_acidity: 0.12, citric_acid: 0.34, residual_sugar: 10.64, chlorides: 0.01, free_sulfur_dioxide: 45.06, total_sulfur_dioxide: 223.43, density: 0.99, pH: 2.95, sulphates: 0.39, alcohol: 8.42}
- {fixed_acidity: 7.85, volatile_acidity: 0.17, citric_acid: 0.47, residual_sugar: 11.68, chlorides: 0.02, free_sulfur_dioxide: 54.28, total_sulfur_dioxide: 181.34, density: 1.0, pH: 2.84, sulphates: 0.26, alcohol: 8.01}
- {fixed_acidity: 7.55, volatile_acidity: 0.59, citric_acid: 0.26, residual_sugar: 16.57, chlorides: 0.01, free_sulfur_dioxide: 19.14, total_sulfur_dioxide: 38.98, density: 1.0, pH: 3.19, sulphates: 0.64, alcohol: 8.09}
- {fixed_acidity: 9.21, volatile_acidity: 0.52, citric_acid: 0.44, residual_sugar: 11.29, chlorides: 0.02, free_sulfur_dioxide: 13.96, total_sulfur_dioxide: 33.78, density: 1.0, pH: 3.03, sulphates: 0.29, alcohol: 8.1}
- {fixed_acidity: 7.84, volatile_acidity: 0.29, citric_acid: 0.23, residual_sugar: 7.12, chlorides: 0.07, free_sulfur_dioxide: 47.37, total_sulfur_dioxide: 72.84, density: 0.99, pH: 3.31, sulphates: 0.87, alcohol: 12.69}
- {fixed_acidity: 8.23, volatile_acidity: 0.45, citric_acid: 0.25, residual_sugar: 11.29, chlorides: 0.02, free_sulfur_dioxide: 26.92, total_sulfur_dioxide: 100.18, density: 1.0, pH: 3.24, sulphates: 0.39, alcohol: 8.17}
- {fixed_acidity: 6.35, volatile_acidity: 0.35, citric_acid: 0.13, residual_sugar: 16.77, chlorides: 0.04, free_sulfur_dioxide: 73.0, total_sulfur_dioxide: 119.71, density: 1.0, pH: 3.02, sulphates: 0.47, alcohol: 10.06}

F.2 MAGIC Gamma

Faithfulness: 0.94

Explanation: The classifiers diverge when encountering samples that exhibit a distinctive pattern characterized by dimensions with lengths ranging from approximately 6.26 to 140.86 and widths from about 0.51 to 63.07, suggesting a compact and slender morphology. These samples typically possess sizes between 1.96 and 3.92, indicating a moderate scale that aligns with their dimensions. The concentration values, both fConc and fConc1, are generally low, with fConc values falling between 0.021 and 0.707 and fConc1 values ranging from 0.002 to 0.376, reflecting a tendency towards lower density characteristics. The asymmetry values, fAsym, vary widely but often fall within the range of -165.53 to 207.44, indicating a diverse range of shapes, while the fM3Long and fM3Trans values suggest a complex internal structure, with fM3Long values typically between -134.53 and 112.33 and fM3Trans values from -87.86 to 60.27. The alpha values, which range from approximately 43.20 to 81.09, further emphasize the unique structural properties of these samples. Additionally, the distances, fDist, span from about 15.62 to 389.29, indicating a variety of spatial distributions. Collectively, these features create a cohesive profile that defines the positive samples, distinguishing them from negative counterparts through their compact size, low density, and diverse asymmetry, making them identifiable in new, unseen data. The classifiers converge when samples do not fit this distinctive pattern.

Divergent Samples:

- {fLength: 25.4, fWidth: 2.56, fSize: 2.94, fConc: 0.09, fConc1: 0.04, fAsym: -89.08, fM3Long: -63.26, fM3Trans: 25.94, fAlpha: 43.2, fDist: 102.61}
- {fLength: 41.23, fWidth: 3.59, fSize: 2.95, fConc: 0.11, fConc1: 0.03, fAsym: -85.98, fM3Long: 47.34, fM3Trans: 34.81, fAlpha: 46.26, fDist: 24.02}
- {fLength: 45.52, fWidth: 9.49, fSize: 3.16, fConc: 0.18, fConc1: 0.03, fAsym: -165.53, fM3Long: -52.43, fM3Trans: 36.74, fAlpha: 53.28, fDist: 220.74}
- {fLength: 13.19, fWidth: 4.36, fSize: 2.99, fConc: 0.19, fConc1: 0.03, fAsym: 65.89, fM3Long: -53.57, fM3Trans: -7.62, fAlpha: 53.64, fDist: 15.62}
- {fLength: 50.8, fWidth: 9.23, fSize: 2.75, fConc: 0.39, fConc1: 0.19, fAsym: -66.35, fM3Long: 8.0, fM3Trans: 0.09, fAlpha: 74.52, fDist: 296.37}
- {fLength: 140.86, fWidth: 26.92, fSize: 3.05, fConc: 0.22, fConc1: 0.13, fAsym: -34.32, fM3Long: 70.71, fM3Trans: 35.97, fAlpha: 58.05, fDist: 203.44}
- {fLength: 41.56, fWidth: 10.51, fSize: 2.85, fConc: 0.02, fConc1: 0.01, fAsym: 161.98, fM3Long: 22.25, fM3Trans: 1.64, fAlpha: 51.75, fDist: 202.45}
- {fLength: 11.21, fWidth: 10.26, fSize: 3.29, fConc: 0.17, fConc1: 0.05, fAsym: 140.28, fM3Long: 112.33, fM3Trans: -1.84, fAlpha: 62.91, fDist: 121.89}
- {fLength: 6.26, fWidth: 2.05, fSize: 1.96, fConc: 0.46, fConc1: 0.31, fAsym: 160.94, fM3Long: -10.24, fM3Trans: 40.21, fAlpha: 76.14, fDist: 19.08}
- {fLength: 22.43, fWidth: 5.9, fSize: 2.51, fConc: 0.13, fConc1: 0.02, fAsym: -19.86, fM3Long: 107.2, fM3Trans: 25.55, fAlpha: 58.41, fDist: 193.56}

Convergent Samples:

- {fLength: 109.19, fWidth: 36.15, fSize: 3.25, fConc: 0.33, fConc1: 0.17, fAsym: -143.84, fM3Long: 16.55, fM3Trans: 10.89, fAlpha: 44.55, fDist: 316.14}
- {fLength: 65.97, fWidth: 34.36, fSize: 2.68, fConc: 0.31, fConc1: 0.12, fAsym: -210.99, fM3Long: -100.89, fM3Trans: 34.42, fAlpha: 22.05, fDist: 350.74}
- {fLength: 94.67, fWidth: 54.87, fSize: 3.29, fConc: 0.04, fConc1: 0.04, fAsym: -221.32, fM3Long: -53.0, fM3Trans: 27.1, fAlpha: 25.47, fDist: 199.49}
- {fLength: 27.71, fWidth: 24.87, fSize: 2.53, fConc: 0.39, fConc1: 0.12, fAsym: 32.83, fM3Long: 88.38, fM3Trans: 23.62, fAlpha: 45.45, fDist: 104.09}

- {fLength: 66.96, fWidth: 28.97, fSize: 3.74, fConc: 0.26, fConc1: 0.14, fAsym: -23.99, fM3Long: 72.42, fM3Trans: 67.6, fAlpha: 24.12, fDist: 283.52}
- {fLength: 113.81, fWidth: 15.64, fSize: 2.88, fConc: 0.29, fConc1: 0.09, fAsym: -171.73, fM3Long: -60.98, fM3Trans: -15.34, fAlpha: 21.6, fDist: 258.8}
- {fLength: 79.83, fWidth: 17.95, fSize: 2.14, fConc: 0.67, fConc1: 0.39, fAsym: 48.33, fM3Long: -89.49, fM3Trans: 66.06, fAlpha: 65.97, fDist: 227.17}
- {fLength: 60.37, fWidth: 33.59, fSize: 3.52, fConc: 0.3, fConc1: 0.11, fAsym: -82.88, fM3Long: -98.04, fM3Trans: -15.34, fAlpha: 84.24, fDist: 201.47}
- {fLength: 108.53, fWidth: 19.74, fSize: 2.67, fConc: 0.55, fConc1: 0.26, fAsym: -217.19, fM3Long: -3.97, fM3Trans: -8.01, fAlpha: 18.36, fDist: 298.34}
- {fLength: 52.78, fWidth: 15.9, fSize: 2.4, fConc: 0.76, fConc1: 0.42, fAsym: -212.02, fM3Long: 26.81, fM3Trans: 7.42, fAlpha: 71.37, fDist: 174.28}

F.3 Rice

Faithfulness: 0.83

Explanation: The classifiers diverge when encountering samples that exhibit a distinctive pattern characterized by a balanced interplay of geometric properties, where the area ranges from approximately 7,800 to 18,400 square units, and the perimeter spans from about 360 to 550 units. These samples consistently demonstrate a major axis length between 145 and 239 units and a minor axis length that typically falls between 64 and 107 units, reflecting a well-defined elliptical shape. The eccentricity values, which range from 0.82 to 0.95, indicate a moderate to high degree of elongation, while the convex area closely aligns with the actual area, suggesting minimal irregularity in shape. Furthermore, the extent values, ranging from 0.49 to 0.86, highlight a strong correlation between the area and the convex area, reinforcing the samples' compactness and structural integrity. This cohesive combination of features not only defines the essence of the positive set but also serves as a reliable framework for classifying new, unseen samples that share these geometric characteristics. The classifiers converge when samples do not exhibit this specific pattern of geometric properties and relationships.

Divergent Samples:

- {Area: 17197.34, Perimeter: 548.45, Major_Axis_Length: 238.92, Minor_Axis_Length: 94.29, Eccentricity: 0.92, Convex_Area: 17551.86, Extent: 0.84}
- {Area: 10107.45, Perimeter: 392.05, Major_Axis_Length: 150.51, Minor_Axis_Length: 86.27, Eccentricity: 0.85, Convex_Area: 10237.1, Extent: 0.73}
- {Area: 18424.43, Perimeter: 548.26, Major_Axis_Length: 227.01, Minor_Axis_Length: 107.49, Eccentricity: 0.88, Convex_Area: 18791.85, Extent: 0.8}
- {Area: 18401.71, Perimeter: 548.26, Major_Axis_Length: 227.67, Minor_Axis_Length: 107.11, Eccentricity: 0.88, Convex_Area: 18746.34, Extent: 0.81}
- {Area: 8903.08, Perimeter: 375.57, Major_Axis_Length: 145.55, Minor_Axis_Length: 80.8, Eccentricity: 0.84, Convex_Area: 9224.63, Extent: 0.72}
- {Area: 10539.21, Perimeter: 440.33, Major_Axis_Length: 195.61, Minor_Axis_Length: 68.03, Eccentricity: 0.95, Convex_Area: 10851.4, Extent: 0.74}
- {Area: 11198.2, Perimeter: 451.88, Major_Axis_Length: 201.98, Minor_Axis_Length: 70.62, Eccentricity: 0.95, Convex_Area: 11590.84, Extent: 0.65}
- {Area: 14902.21, Perimeter: 471.76, Major_Axis_Length: 183.14, Minor_Axis_Length: 104.28, Eccentricity: 0.82, Convex_Area: 15310.79, Extent: 0.7}
- {Area: 17345.04, Perimeter: 524.78, Major_Axis_Length: 211.36, Minor_Axis_Length: 107.49, Eccentricity: 0.86, Convex_Area: 17665.62, Extent: 0.81}
- {Area: 17708.63, Perimeter: 536.71, Major_Axis_Length: 218.11, Minor_Axis_Length: 106.87, Eccentricity: 0.87, Convex_Area: 18086.54, Extent: 0.8}

Convergent Samples:

- {Area: 12345.76, Perimeter: 418.18, Major_Axis_Length: 158.2, Minor_Axis_Length: 97.84, Eccentricity: 0.81, Convex_Area: 12591.93, Extent: 0.74}
- {Area: 15686.19, Perimeter: 495.05, Major_Axis_Length: 197.67, Minor_Axis_Length: 103.32, Eccentricity: 0.85, Convex_Area: 15936.47, Extent: 0.67}
- {Area: 16776.94, Perimeter: 533.87, Major_Axis_Length: 222.79, Minor_Axis_Length: 98.47, Eccentricity: 0.89, Convex_Area: 17039.94, Extent: 0.65}
- {Area: 16913.29, Perimeter: 545.04, Major_Axis_Length: 237.79, Minor_Axis_Length: 91.17, Eccentricity: 0.94, Convex_Area: 17062.7, Extent: 0.63}
- {Area: 17401.85, Perimeter: 547.88, Major_Axis_Length: 238.92, Minor_Axis_Length: 95.2, Eccentricity: 0.93, Convex_Area: 17574.62, Extent: 0.68}
- {Area: 15936.16, Perimeter: 533.68, Major_Axis_Length: 234.32, Minor_Axis_Length: 87.52, Eccentricity: 0.95, Convex_Area: 16141.24, Extent: 0.63}
- {Area: 10107.45, Perimeter: 381.82, Major_Axis_Length: 148.26, Minor_Axis_Length: 87.38, Eccentricity: 0.83, Convex_Area: 10373.61, Extent: 0.81}
- {Area: 16288.38, Perimeter: 543.33, Major_Axis_Length: 237.98, Minor_Axis_Length: 88.48, Eccentricity: 0.95, Convex_Area: 16653.16, Extent: 0.66}
- {Area: 9368.92, Perimeter: 375.57, Major_Axis_Length: 147.42, Minor_Axis_Length: 83.11, Eccentricity: 0.84, Convex_Area: 9725.18, Extent: 0.74}
- {Area: 16072.5, Perimeter: 537.65, Major_Axis_Length: 233.85, Minor_Axis_Length: 87.14, Eccentricity: 0.95, Convex_Area: 16562.15, Extent: 0.63}
- {Area: 11323.18, Perimeter: 450.55, Major_Axis_Length: 201.51, Minor_Axis_Length: 71.1, Eccentricity: 0.95, Convex_Area: 11568.09, Extent: 0.53}
- {Area: 16072.5, Perimeter: 539.74, Major_Axis_Length: 234.79, Minor_Axis_Length: 86.47, Eccentricity: 0.95, Convex_Area: 16505.27, Extent: 0.53}
- {Area: 15458.95, Perimeter: 523.83, Major_Axis_Length: 230.95, Minor_Axis_Length: 85.12, Eccentricity: 0.95, Convex_Area: 15572.44, Extent: 0.63}
- {Area: 10982.32, Perimeter: 390.72, Major_Axis_Length: 145.26, Minor_Axis_Length: 95.11, Eccentricity: 0.81, Convex_Area: 11306.44, Extent: 0.72}
- {Area: 18322.18, Perimeter: 544.85, Major_Axis_Length: 225.23, Minor_Axis_Length: 107.45, Eccentricity: 0.87, Convex_Area: 18689.46, Extent: 0.65}

F.4 Bank Note

Faithfulness: 0.84

Divergent Samples:

- variance: -1.43, skewness: 5.23, curtosis: 0.94, entropy: -4.44
- variance: -3.28, skewness: 2.96, curtosis: -2.52, entropy: -1.84
- variance: -1.54, skewness: 1.22, curtosis: -2.5, entropy: -3.95
- variance: -2.06, skewness: 9.58, curtosis: -3.45, entropy: -1.2
- variance: -6.89, skewness: 2.8, curtosis: -0.06, entropy: -3.96
- variance: -3.09, skewness: 4.03, curtosis: 2.33, entropy: -5.03
- variance: -4.67, skewness: 5.34, curtosis: -0.6, entropy: -0.33
- variance: -1.37, skewness: 7.18, curtosis: -1.32, entropy: -0.87
- variance: 1.07, skewness: 3.6, curtosis: -2.9, entropy: -3.91
- variance: -3.13, skewness: 6.27, curtosis: -0.39, entropy: -6.34

Convergent Samples:

- variance: -1.43, skewness: 5.23, curtosis: 0.94, entropy: -4.44
- variance: -3.28, skewness: 2.96, curtosis: -2.52, entropy: -1.84
- variance: -1.54, skewness: 1.22, curtosis: -2.5, entropy: -3.95
- variance: -2.06, skewness: 9.58, curtosis: -3.45, entropy: -1.2
- variance: -6.89, skewness: 2.8, curtosis: -0.06, entropy: -3.96
- variance: -3.09, skewness: 4.03, curtosis: 2.33, entropy: -5.03
- variance: -4.67, skewness: 5.34, curtosis: -0.6, entropy: -0.33
- variance: -1.37, skewness: 7.18, curtosis: -1.32, entropy: -0.87
- variance: 1.07, skewness: 3.6, curtosis: -2.9, entropy: -3.91
- variance: -3.13, skewness: 6.27, curtosis: -0.39, entropy: -6.34

F.5 Adult

Faithfulness: 0.94

Explanation: The classifiers diverge when examining individuals predominantly aged between 21 and 74 years, with a notable concentration around the mid-40s to early 50s, reflecting a diverse range of life experiences. Their education levels vary, with education-num values typically ranging from 5.44 to 15.95, indicating a mix of lower to higher educational attainment, often associated with occupations in skilled trades, management, and service sectors. Capital gains are generally modest, falling between 0 and 18699, while capital losses remain low, typically under 900, suggesting a stable financial situation without extreme fluctuations. The hours worked per week vary from approximately 12.76 to 59.51, with many individuals working around 25 to 45 hours, indicating a balance between work and personal life. Marital statuses are diverse, with many individuals being married or separated, and occupations span a range of fields, including craft-repair, managerial roles, and service positions. This combination of moderate age, varied education, stable financial indicators, and diverse work hours and marital statuses creates a distinct profile that sets these samples apart from the negative set, making them identifiable as part of the positive group. The classifiers converge when these characteristics are not present, indicating a different pattern in the negative samples.

Divergent Samples:

- age: 45.32, fnlwgt: 124621.74, education-num: 9.87, capital-gain: 3799.96, capital-loss: 418.18, hours-per-week: 25.3, workclass: ?, education: 9th, marital-status: Married-spouse-absent, occupation: Craft-repair, relationship: Husband, race: White, sex: Male, native-country: Laos
- age: 47.08, fnlwgt: 61062.79, education-num: 9.07, capital-gain: 1799.98, capital-loss: 78.41, hours-per-week: 28.54, workclass: Local-gov, education: HS-grad, marital-status: Married-spouse-absent, occupation: Priv-house-serv, relationship: Wife, race: White, sex: Male, native-country: Haiti
- age: 35.83, fnlwgt: 200005.61, education-num: 11.92, capital-gain: 18699.81, capital-loss: 148.1, hours-per-week: 38.24, workclass: ?, education: Assoc-acdm, marital-status: Married-civ-spouse, occupation: Exec-managerial, relationship: Wife, race: Black, sex: Male, native-country: Guatemala
- age: 48.54, fnlwgt: 157140.27, education-num: 13.35, capital-gain: 7499.93, capital-loss: 348.48, hours-per-week: 45.79, workclass: Never-worked, education: 11th, marital-status: Divorced, occupation: Transport-moving, relationship: Husband, race: Asian-Pac-Islander, sex: Female, native-country: Portugal
- age: 35.91, fnlwgt: 32978.61, education-num: 6.45, capital-gain: 300.0, capital-loss: 13.07, hours-per-week: 28.44, workclass: Local-gov, education: Bachelors, marital-status: Separated, occupation: Other-service, relationship: Husband, race: White, sex: Female, native-country: Guatemala
- age: 47.66, fnlwgt: 251739.63, education-num: 11.89, capital-gain: 5799.94, capital-loss: 823.28, hours-per-week: 28.54, workclass: Never-worked, education: HS-grad,

marital-status: Married-spouse-absent, occupation: Handlers-cleaners, relationship: Husband, race: Black, sex: Female, native-country: Germany

- age: 38.68, fnlwgt: 24109.92, education-num: 5.44, capital-gain: 200.0, capital-loss: 8.71, hours-per-week: 31.97, workclass: Local-gov, education: HS-grad, marital-status: Divorced, occupation: Other-service, relationship: Not-in-family, race: Black, sex: Male, native-country: Japan
- age: 49.48, fnlwgt: 31500.5, education-num: 15.4, capital-gain: 499.99, capital-loss: 34.85, hours-per-week: 12.76, workclass: Local-gov, education: 11th, marital-status: Separated, occupation: Craft-repair, relationship: Wife, race: White, sex: Male, native-country: Portugal
- age: 58.17, fnlwgt: 35934.84, education-num: 15.44, capital-gain: 3199.97, capital-loss: 43.56, hours-per-week: 40.49, workclass: Self-emp-inc, education: Masters, marital-status: Married-spouse-absent, occupation: Exec-managerial, relationship: Not-in-family, race: Asian-Pac-Islander, sex: Male, native-country: Haiti
- age: 68.03, fnlwgt: 15241.23, education-num: 8.2, capital-gain: 200.0, capital-loss: 4.36, hours-per-week: 38.83, workclass: Federal-gov, education: 9th, marital-status: Separated, occupation: Handlers-cleaners, relationship: Not-in-family, race: White, sex: Male, native-country: Guatemala

Convergent Samples:

- age: 50.8, fnlwgt: 137924.78, education-num: 11.84, capital-gain: 8699.91, capital-loss: 892.98, hours-per-week: 30.11, workclass: ?, education: 5th-6th, marital-status: Divorced, occupation: Craft-repair, relationship: Wife, race: Black, sex: Male, native-country: El-Salvador
- age: 39.19, fnlwgt: 155662.16, education-num: 10.92, capital-gain: 13099.87, capital-loss: 1507.18, hours-per-week: 37.55, workclass: Self-emp-not-inc, education: 7th-8th, marital-status: Separated, occupation: Prof-specialty, relationship: Wife, race: Black, sex: Female, native-country: Poland
- age: 45.1, fnlwgt: 117231.17, education-num: 9.28, capital-gain: 7999.92, capital-loss: 317.99, hours-per-week: 39.61, workclass: Self-emp-inc, education: 9th, marital-status: Separated, occupation: Handlers-cleaners, relationship: Wife, race: White, sex: Male, native-country: Cuba
- age: 39.12, fnlwgt: 353729.57, education-num: 10.78, capital-gain: 4799.95, capital-loss: 339.77, hours-per-week: 38.04, workclass: Self-emp-inc, education: 11th, marital-status: Never-married, occupation: Prof-specialty, relationship: Other-relative, race: Asian-Pac-Islander, sex: Female, native-country: Germany
- age: 65.55, fnlwgt: 160096.5, education-num: 8.51, capital-gain: 16099.84, capital-loss: 400.75, hours-per-week: 59.31, workclass: Federal-gov, education: Assoc-voc, marital-status: Married-spouse-absent, occupation: Tech-support, relationship: Other-relative, race: Asian-Pac-Islander, sex: Male, native-country: Philippines
- age: 57.0, fnlwgt: 469022.51, education-num: 8.62, capital-gain: 4699.95, capital-loss: 378.97, hours-per-week: 34.12, workclass: ?, education: Assoc-acdm, marital-status: Married-spouse-absent, occupation: ?, relationship: Wife, race: Black, sex: Male, native-country: Germany
- age: 50.95, fnlwgt: 599096.68, education-num: 12.79, capital-gain: 3899.96, capital-loss: 553.21, hours-per-week: 34.81, workclass: ?, education: HS-grad, marital-status: Separated, occupation: Prof-specialty, relationship: Husband, race: Black, sex: Female, native-country: Puerto-Rico
- age: 55.62, fnlwgt: 157140.27, education-num: 14.33, capital-gain: 13499.87, capital-loss: 291.85, hours-per-week: 48.33, workclass: State-gov, education: 9th, marital-status: Separated, occupation: Sales, relationship: Wife, race: Black, sex: Male, native-country: Canada
- age: 45.18, fnlwgt: 112796.83, education-num: 9.13, capital-gain: 4599.95, capital-loss: 853.78, hours-per-week: 50.39, workclass: Federal-gov, education: Masters,

marital-status: Married-AF-spouse, occupation: Machine-op-inspct, relationship: Wife, race: Black, sex: Male, native-country: Honduras

- age: 47.0, fnlwgt: 142359.12, education-num: 12.1, capital-gain: 14099.86, capital-loss: 483.52, hours-per-week: 34.32, workclass: ?, education: Assoc-voc, marital-status: Separated, occupation: Farming-fishing, relationship: Not-in-family, race: White, sex: Male, native-country: Honduras

F.6 Bank Marketing

Faithfulness: 0.68

Explanation: The classifiers diverge when individuals are predominantly in the age range of 18 to 54 years, with a notable tendency towards younger adults, particularly those aged between 21 and 39 years. These individuals typically exhibit a negative balance, ranging from approximately -3172 to -7908, indicating financial challenges. The duration of their interactions tends to vary significantly, spanning from brief engagements of around 0 to 73 seconds, with many samples reflecting durations between 19 and 49 seconds. The campaign variable shows a range from 1.0 to 4.78, suggesting a moderate level of outreach efforts. Most samples are associated with a marital status of 'married' and possess a tertiary education, which appears to correlate with their job roles, predominantly in sectors such as management and blue-collar work. The contact method is primarily cellular, and the outcome of their previous interactions is consistently successful, reinforcing a pattern of resilience despite financial difficulties. This combination of age, financial status, interaction duration, and educational background creates a distinct profile. The classifiers converge when these characteristics are not present.

Divergent Samples:

- age: 39.18, balance: -7468.27, day: 4.36, month: jun, duration: 9.84, campaign: 1.06, pdays: 164.68, previous: 0.28, job: entrepreneur, marital: married, education: tertiary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success

- age: 47.64, balance: -3172.58, day: 18.01, month: may, duration: 162.29, campaign: 2.43, pdays: 335.59, previous: 5.77, job: management, marital: married, education: tertiary, default: no, housing: no, loan: no, contact: cellular, poutcome: success

- age: 29.32, balance: -6256.66, day: 7.69, month: may, duration: 29.51, campaign: 1.37, pdays: 435.0, previous: 1.93, job: blue-collar, marital: married, education: tertiary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success

- age: 26.24, balance: -6366.81, day: 8.26, month: may, duration: 19.67, campaign: 1.31, pdays: 337.34, previous: 1.1, job: blue-collar, marital: married, education: tertiary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success

- age: 25.85, balance: -6697.25, day: 5.71, month: may, duration: 34.43, campaign: 1.43, pdays: 441.98, previous: 2.2, job: entrepreneur, marital: married, education: tertiary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success

- age: 29.7, balance: -5375.5, day: 22.0, month: may, duration: 359.01, campaign: 4.78, pdays: 331.23, previous: 4.12, job: management, marital: married, education: tertiary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success

- age: 42.41, balance: -4274.04, day: 13.45, month: may, duration: 270.49, campaign: 3.23, pdays: 338.21, previous: 7.7, job: management, marital: married, education: tertiary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success

- age: 39.1, balance: 5749.25, day: 3.13, month: jun, duration: 34.43, campaign: 1.37, pdays: 180.38, previous: 1.1, job: entrepreneur, marital: married, education: tertiary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success

- age: 21.77, balance: -6366.81, day: 4.54, month: may, duration: 290.16, campaign: 1.87, pdays: 177.76, previous: 1.37, job: nan, marital: married, education: tertiary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success

- age: 39.56, balance: -6807.39, day: 14.83, month: may, duration: 63.93, campaign: 1.12, pdays: 140.26, previous: 0.83, job: management, marital: married, education: tertiary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success

Convergent Samples:

- age: 49.72, balance: -2181.26, day: 11.71, month: may, duration: 73.77, campaign: 5.15, pdays: 583.24, previous: 11.28, job: retired, marital: single, education: secondary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success
- age: 39.18, balance: -1189.95, day: 8.59, month: may, duration: 118.03, campaign: 1.87, pdays: 491.68, previous: 7.15, job: technician, marital: single, education: secondary, default: no, housing: no, loan: no, contact: cellular, poutcome: other
- age: 31.01, balance: -2291.41, day: 12.49, month: may, duration: 250.82, campaign: 2.67, pdays: 420.18, previous: 7.43, job: entrepreneur, marital: single, education: secondary, default: no, housing: no, loan: yes, contact: cellular, poutcome: success
- age: 25.39, balance: -7247.98, day: 15.94, month: may, duration: 19.67, campaign: 1.25, pdays: 491.68, previous: 0.83, job: entrepreneur, marital: married, education: secondary, default: no, housing: yes, loan: no, contact: telephone, poutcome: success
- age: 35.86, balance: -4824.77, day: 20.05, month: may, duration: 93.44, campaign: 4.97, pdays: 626.84, previous: 7.15, job: technician, marital: single, education: nan, default: no, housing: yes, loan: no, contact: cellular, poutcome: success
- age: 31.86, balance: -2731.99, day: 8.89, month: may, duration: 127.87, campaign: 1.81, pdays: 402.74, previous: 6.6, job: entrepreneur, marital: single, education: secondary, default: no, housing: no, loan: no, contact: cellular, poutcome: success
- age: 36.4, balance: -6476.96, day: 24.97, month: may, duration: 39.34, campaign: 4.29, pdays: 513.48, previous: 2.75, job: management, marital: single, education: secondary, default: no, housing: yes, loan: no, contact: telephone, poutcome: success
- age: 35.63, balance: -5155.2, day: 13.06, month: may, duration: 216.39, campaign: 1.93, pdays: 233.57, previous: 6.05, job: technician, marital: single, education: secondary, default: no, housing: yes, loan: no, contact: telephone, poutcome: other
- age: 35.56, balance: -749.36, day: 15.04, month: may, duration: 368.85, campaign: 3.17, pdays: 400.99, previous: 6.6, job: admin., marital: divorced, education: secondary, default: no, housing: yes, loan: no, contact: cellular, poutcome: success
- age: 55.27, balance: 7952.17, day: 19.03, month: may, duration: 147.54, campaign: 3.36, pdays: 205.66, previous: 3.58, job: technician, marital: married, education: nan, default: no, housing: no, loan: no, contact: cellular, poutcome: success

F.7 Balance Scale

Faithfulness: 0.52

Explanation: The classifiers diverge when there is a harmonious balance between right and left distances and weights, where right distances predominantly range from 1 to 5 and right weights vary from 1 to 5, often reflecting a tendency for higher values in both dimensions. Specifically, the right distance frequently reaches values of 4 or 5, while the left distance typically spans from 1 to 5, with many instances showcasing a left weight that also aligns closely with the right weight, often falling between 1 and 5. This interplay creates a cohesive structure where the right and left attributes complement each other, resulting in configurations that exhibit a notable symmetry or proportionality, particularly when both right and left weights are elevated. The instances also demonstrate a tendency for the left distance to be equal to or greater than the right distance, fostering a sense of equilibrium. The classifiers converge when this distinctive pattern of balanced proportions and complementary values is absent, making the instances easily identifiable and classifiable based on their intrinsic relationships among the features.

Divergent Samples:

- right-distance: 4, right-weight: 3, left-distance: 5, left-weight: 3

- right-distance: 4, right-weight: 1, left-distance: 1, left-weight: 4
- right-distance: 5, right-weight: 5, left-distance: 5, left-weight: 5
- right-distance: 5, right-weight: 2, left-distance: 3, left-weight: 3
- right-distance: 5, right-weight: 3, left-distance: 3, left-weight: 5
- right-distance: 5, right-weight: 2, left-distance: 4, left-weight: 5
- right-distance: 5, right-weight: 4, left-distance: 2, left-weight: 5
- right-distance: 3, right-weight: 2, left-distance: 2, left-weight: 2
- right-distance: 1, right-weight: 5, left-distance: 4, left-weight: 5
- right-distance: 5, right-weight: 1, left-distance: 2, left-weight: 1

Convergent Samples:

- right-distance: 2, right-weight: 3, left-distance: 5, left-weight: 1
- right-distance: 1, right-weight: 1, left-distance: 3, left-weight: 2
- right-distance: 3, right-weight: 3, left-distance: 5, left-weight: 2
- right-distance: 1, right-weight: 1, left-distance: 2, left-weight: 1
- right-distance: 1, right-weight: 1, left-distance: 5, left-weight: 2
- right-distance: 3, right-weight: 3, left-distance: 5, left-weight: 2
- right-distance: 1, right-weight: 1, left-distance: 3, left-weight: 2
- right-distance: 3, right-weight: 3, left-distance: 5, left-weight: 2
- right-distance: 1, right-weight: 1, left-distance: 5, left-weight: 3
- right-distance: 1, right-weight: 1, left-distance: 5, left-weight: 3