

MIRAGE: A Model-Inference Framework for Rule-Based and Grounded Explanations of Black-Box Neural Systems

Dhruv Patel, Soham Gupta

Department of Computer Science, Harvard University, Cambridge, MA, USA
dhruvpatel@college.harvard.edu, sohamgupta@college.harvard.edu

Abstract

While machine learning models continue to grow in both accuracy and real-world impact, progress in their interpretability has lagged behind their capabilities. In many settings, highly effective models are deployed as black boxes whose internal mechanisms are inaccessible, making it difficult to reason about their behavior. To address the issue of black box models which are internally inaccessible, we introduce MIRAGE (Model Inference for Rule-Based and Grounded Explanations), a framework built to generate logical theories to explain black-box models without access to internal representations. MIRAGE observes inputs and predictions and induces Horn clause rules over a set of feature predicates. Unlike local explanation approaches, MIRAGE generates a theory that gives stable decision patterns across inputs. We find that on synthetic datasets, MIRAGE works almost without fault, resulting in high Global Explainability Index (GEI) scores of approximately 0.88. This reflects complete coverage, low rule complexity, as well as high logical consistency. When tested on real-world datasets, with serious levels of heterogeneity, the framework achieves GEI scores between 0.64 and 0.83. MIRAGE then demonstrates that classical inference techniques, when applied to modern ML systems, are capable of yielding effective explanations for black-box models.

1 Introduction and Motivation

Modern machine learning models are effective, but they’re often difficult to understand or reason about in a concrete and grounded way. In many real-world systems, as we’ve learned over this course, models aren’t used in isolation and rather serve as one component of larger decision-making pipelines. In compositional AI, this opacity becomes a serious limitation. After all, if the behavior of an individual model cannot be clearly characterized, it will be even more difficult to reason about how it will interact with other components or to verify the behavior of the system as a whole.

Symbolic representations should then offer a natural way to address this issue and support compositional reasoning at scale. In particular, logical rules are uniquely capable of making decision processes explicit and modular, allowing them to most importantly be reused. That said, though, the models that are the most accurate today typically operate in continuous, high-dimensional spaces and do not expose such structure directly — and if

they do, still mask it through other mechanisms. This creates a gap between models that perform well and representations that are easy to reason about compositionally.

Explainability methods attempt to bridge this gap, but many focus on explaining individual predictions rather than capturing a model’s behavior at a global level. While these explanations can be useful in isolation, they do not provide a clear picture of how a model behaves across its entire input space. Because of this, they offer very little support for tasks like system-level analysis or integration with other symbolic components that require reasoning at a global level.

In this work, we explore whether a trained black-box classifier, like a NN, can be summarized by a small set of symbolic rules that reflect its decision logic at a global level. Instead of attempting to modify the model or inspect its internal parameters, as much existing literature in this space has focused on for years, we treat the model as a black box and focus only on its observable input–output behavior. The focus should then be on producing a symbolic description that is both faithful to the model and structured in a way that supports compositional reasoning.

To that end, we introduce MIRAGE, a framework that induces logical rules from a trained model’s predictions. MIRAGE incrementally builds a global rule set, refines it to improve coverage and consistency, and resolves contradictions as they arise based on specific metrics. The result of this is a single symbolic theory that aims to capture how the model behaves, rather than why it made a specific prediction in isolation.

MIRAGE then hopes to answer a straightforward, yet increasingly important question: can the behavior of a black-box classifier be captured as a compact, compositional set of symbolic rules without significantly departing from the model’s original predictions? In particular, we study this question by evaluating MIRAGE across multiple datasets and model classes, measuring how well the induced rules approximate the model while remaining concise / interpretable.

2 Related Work

As it stands today, there are two main approaches in explainable artificial intelligence — transparent and post-hoc approaches. Transparent models (e.g., decision trees or rule-based classifiers) are designed to be interpretable from the outset, meaning that they should be exposing their decision logic directly to the user. However, these models often struggle to match the performance of more expressive learners on complex tasks. Post-hoc methods, on the other hand, seek to explain the behavior of an already trained black-box model, attempting to balance the accuracy of predictions and how clearly it’s able to explain how decisions are made.

Many post-hoc methods focus on local explanations, which try to justify individual predictions. Techniques such as LIME, SHAP, and Anchors explain model outputs by identifying influential features or sufficient conditions in a small neighborhood around a given input (Lundberg & Lee, 2017; Ribeiro et al., 2016, 2018). These methods are ultimately model-agnostic and meant to be intuitive, making them useful for assessing trust in specific predictions. However, because they are local by design, they do not provide a single expla-

nation of a model’s logic at a global level. Understanding overall behavior would require aggregating many local explanations, which can quickly become computationally expensive and just conceptually unclear.

To address these limitations, prior work has turned to global surrogate models that approximate a black-box classifier across the full input space. Early approaches treated trained neural networks as oracles and induced decision trees that mimicked their predictions (Craven & Shavlik, 1995). Later methods extended this idea to other model classes, including rule extraction for SVMs and distillation techniques that train simpler models on the outputs of complex ones (Fung et al., 2005; Hinton et al., 2015). Though these approaches are (modestly) effective, they often result in super large or deeply nested structures that are difficult to interpret.

Now, in a different vein, symbolic learning methods offer an alternative by representing explanations as logical rules rather than trees or feature-weight vectors. A foundational example is Shapiro’s model inference framework, which we saw in class. It introduced an oracle-based procedure for incrementally learning Horn-clause theories consistent with observed data (Shapiro, 1981). In the framework as defined by the paper, the learner interacts with an oracle that answers queries about ground facts, thereby constructing a logically consistent global theory. Although originally developed in a purely symbolic context, this perspective naturally extends to black-box machine learning models, which to an extent proposes similar input-output oracles. Shapiro’s paper then serves as the driving thrust behind MIRAGE.

Other work has focused on extracting rules directly from deep NNs by trying to capture / approximate internal representations with decision trees and merging them into a global rule set (Zilke et al., 2016). While these methods achieve strong fidelity, they hinge on needing access to the internals of models and are tied to specific architectures. More recent approaches have tried to address those shortcomings, producing compact global rule sets by merging local explanations or solving optimization problems that balance fidelity and simplicity (Dash et al., 2023; Setzu et al., 2021). Though these methods yield concise explanations, they often rely on heuristics that make the resulting logic harder to analyze compositionally and conceptually.

Altogether, this points to a persistent gap between local explainability and globally interpretable explanations. While many methods explain individual predictions or approximate models at a high level, few aim to induce a single, coherent symbolic theory that captures a model’s behavior as a whole without relying on internal model access. MIRAGE tries to fix this, and we hypothesize that it can:

- Induce compact, logically consistent symbolic theories that capture stable decision patterns of black-box models globally; and
- Operate robustly across diverse model classes, producing explanations of comparable quality regardless of architecture.

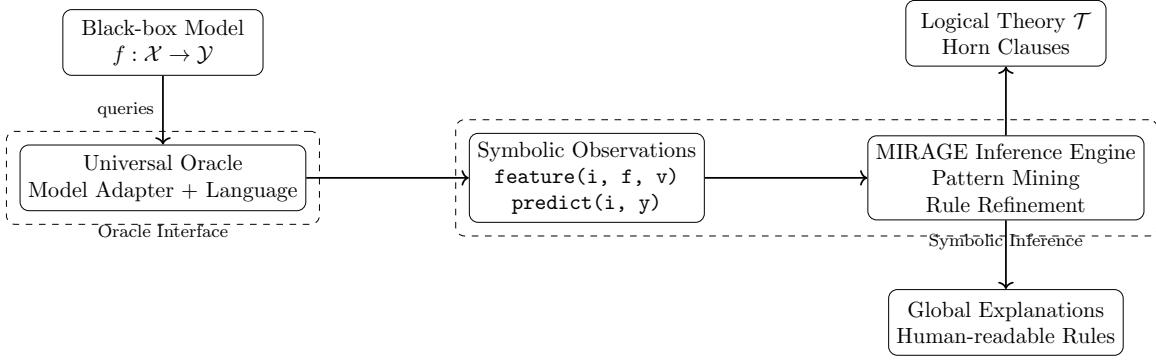


Figure 1: The MIRAGE framework. MIRAGE treats a trained model as a black-box oracle, encodes model behavior as symbolic observations, and incrementally induces a global logical theory composed of Horn clauses.

3 The MIRAGE Framework

3.1 Overview and Architecture

At a high level, MIRAGE operates by observing how a trained model behaves on concrete inputs and incrementally constructing a set of logical rules that approximate the model’s decision function. Figure 1 provides an overview of the system architecture.

Given a trained black-box classifier

$$f : \mathcal{X} \rightarrow \mathcal{Y},$$

MIRAGE does not assume access to internal representations, which can include things like weights, gradients, or hidden activations. Instead, it queries the model only through input-output interactions. These interactions are mediated by a universal oracle, which converts model behavior into symbolic observations suitable for logical inference.

The core output of MIRAGE is a symbolic theory

$$\mathcal{T} = \{C_1, \dots, C_k\},$$

where each clause C_i is a Horn clause describing a sufficient condition under which the model predicts a given label. This theory serves as a human-readable approximation of the model’s behavior.

3.2 Symbolic Representation and Oracle Interface

MIRAGE represents model behavior using a first-order logic observation language. For tabular classification tasks, each instance x_i is encoded as a set of ground feature facts of the form

$$\text{feature}(i, f_j, v_j),$$

where i is an instance identifier, f_j is a feature name, and v_j is a discretized feature value (e.g., `low`, `medium`, `high`).¹

¹Continuous features are discretized using quantile-based binning to balance expressivity and rule simplicity.

Model predictions are represented using prediction facts:

$$\text{predict}(i, y),$$

where $y \in \mathcal{Y}$ is the label predicted by the black-box model for instance i .

The universal oracle is responsible for:

1. Adapting heterogeneous models (for instance, architectures like neural networks, decision trees, SVMs, etc) to a common prediction interface.
2. Generating symbolic facts from raw inputs and model outputs.
3. Answering logical queries during the rule evaluation process.

Formally, the oracle defines a mapping

$$\mathcal{O} : \mathcal{X} \rightarrow \mathcal{F} \cup \mathcal{P},$$

where \mathcal{F} is the set of feature facts and \mathcal{P} is the set of prediction facts. This design ensures that MIRAGE remains fully model-agnostic and applicable across a wide range of learning algorithms.

3.3 Incremental Rule Inference

MIRAGE builds on Shapiro’s incremental Model Inference Algorithm to induce a logical theory from observed facts (Shapiro, 1981). The inference process proceeds in three main phases.

3.3.1 Pattern-Based Rule Generation

Rather than directly inducing a single hypothesis, MIRAGE first enumerates a set of candidate rules derived from observed symbolic facts. Candidate rules are generated by considering conjunctions of feature predicates that co-occur with a given predicted label.

Concretely, for each label $y \in \mathcal{Y}$ and each instance i such that $f(x_i) = y$, MIRAGE constructs candidate bodies by selecting subsets of the instance’s feature facts:

$$\{\text{feature}(i, f_j, v_j)\}_{j \in S}, \quad S \subseteq \{1, \dots, d\}.$$

To control combinatorial growth, MIRAGE prioritizes single-feature candidates and expands to multi-feature conjunctions only when simpler rules fail to achieve sufficient discriminative power.

Each candidate rule r is then evaluated against the full set of observed instances using empirical support and confidence:

$$\text{support}(r) = \frac{|\{i : r \text{ covers } i \wedge f(x_i) = y\}|}{|\{i : f(x_i) = y\}|},$$

$$\text{confidence}(r) = \frac{|\{i : r \text{ covers } i \wedge f(x_i) = y\}|}{|\{i : r \text{ covers } i\}|}.$$

Candidates failing to exceed the predefined thresholds are disregarded and thrown out of the evaluation pool. In the vast majority of cases, and unless otherwise defined, MIRAGE uses modest default thresholds (e.g., minimum support on the order of 10% and minimum confidence above 60%), which were found to balance rule stability and interpretability across datasets.² Among the remaining candidates, MIRAGE ranks candidate rules using a composite score

$$\text{score}(r) = \alpha \cdot \text{confidence}(r) + \beta \cdot \text{support}(r) - \gamma \cdot |r|,$$

where $|r|$ denotes the number of conditions in the rule body, and $\alpha, \beta, \gamma > 0$ control the trade-off between rule generality and interpretability.

3.3.2 Clause Construction

Each accepted rule is encoded as a Horn clause of the form:

$$\text{predict}(X, y) \leftarrow \text{feature}(X, f_1, v_1), \dots, \text{feature}(X, f_m, v_m),$$

where X is a logical variable ranging over instances. We define it as such to enable compact explanations while preserving logical clarity.

3.3.3 Proof Trees and Resolution

Once candidate clauses are introduced into the theory, MIRAGE evaluates their logical consequences using a resolution-based inference engine. For a given instance i and predicted label y , a clause C is said to explain the prediction if $\text{predict}(i, y)$ can be derived from the clause body and observed feature facts.

Component	Description
Instance	i
Observed facts	$\text{feature}(i, \text{income}, \text{low})$ $\text{feature}(i, \text{credit_score}, \text{medium})$
Rule	$\text{predict}(X, \text{DENIED}) \leftarrow \text{feature}(X, \text{income}, \text{low})$
Derived conclusion	$\text{predict}(i, \text{DENIED})$

Table 1: Example proof tree instantiated as a resolution derivation. Ground feature facts satisfy the body of a Horn clause, allowing the prediction atom to be derived for a specific instance.

This derivation is represented as a proof tree, whose leaves correspond to ground feature facts and whose root corresponds to the prediction atom. Each internal node represents an application of a Horn clause via resolution. Figure 1 conceptually illustrates this inference flow. Formally, a proof tree corresponds to a successful SLD-resolution derivation of the goal atom $\text{predict}(i, y)$ from the induced theory and the set of observed ground facts.

²These values were selected empirically by checking support and confidence thresholds over a coarse grid on a held-out subset of instances and choosing ranges that consistently produced stable rule sets with broad applicability, while avoiding excessive rule fragmentation or overfitting.

We find that proof trees serve two purposes. First, they provide a formal justification for why a rule applies to a given instance. Second, they enable MIRAGE to detect contradictions. For instance, let’s say if multiple proof trees derive conflicting predictions for the same instance. If that is the case, MIRAGE would mark the corresponding clauses for refinement or removal.

3.3.4 Theory Refinement

To improve coverage and reduce contradictions, MIRAGE incrementally refines the theory by:

1. Identifying instances not covered by any existing rule for their predicted label.
2. Generating specialized clauses tailored to these uncovered instances.
3. Pruning rules that introduce too many false positives.

This process continues until either full coverage is achieved or a maximum theory size is reached. The final theory, then, balances explanatory breadth with interpretability, meaning that it attempts to produce concise explanations that remain grounded in observed behavior.

In practice, MIRAGE yields theories consisting of approximately 10–12 rules with an average rule length of 1.4–1.5 conditions. This design explicitly favors shallow, human-comprehensible explanations over exhaustive symbolic completeness, reflecting the primary goal of explainability rather than exact logical reconstruction.³

4 Experimental Setup

4.1 Datasets

We evaluate MIRAGE on both synthetic and real-world classification datasets in order to assess its behavior under controlled logical structure as well as practical, noisy settings.

4.1.1 Synthetic Datasets

We consider the following synthetic datasets ⁴:

- **Synthetic Loan:** A tabular dataset with four features (`income`, `credit_score`, `age`, `employment_years`) discretized into categorical bins. Labels are generated using a human-interpretable rule set mimicking credit approval logic.
- **Binary Rule:** A low-dimensional dataset with binary features constructed from simple disjunctive and conjunctive rules.

³This trade-off aligns with prior work emphasizing global interpretability over perfect symbolic equivalence.

⁴We acknowledge the downsides of using synthetic data. But we find that it enables controlled experiments, allowing us to directly assess whether MIRAGE recovers the intended logical structure rather than spurious correlations.

- XOR: A nonlinear classification dataset requiring multi-feature interaction, designed to test MIRAGE’s ability to recover compositional logic.

All synthetic datasets contain approximately 600 instances. We also ensured that we maintained a balanced class distribution.

4.1.2 Real-World Datasets

We additionally evaluate MIRAGE on real-world datasets drawn from standard machine learning benchmarks. Specifically, we use:

- Adult Income: Census data predicting income category from demographic and occupational features.
- Breast Cancer Wisconsin: Diagnostic measurements for malignant vs. benign tumors.

Continuous features are standardized and discretized using quantile-based binning prior to symbolic encoding. We also one-hot code categorical variables before discretization. All in all, Table 2 summarizes the datasets used.

Dataset	#Instances	#Features	Type
Synthetic Loan	~ 600	4	Synthetic
Binary Rule	~ 600	3	Synthetic
XOR	~ 600	2	Synthetic
Adult Income	~ 48,000	~ 100	Real-world
Breast Cancer	569	30	Real-world

Table 2: Datasets used in evaluation. Feature counts reflect post-processing prior to symbolic encoding.

4.2 Models

To evaluate MIRAGE across diverse hypothesis classes, we test four standard classifiers commonly used in explainability studies:

- Neural Network (NN): A multilayer perceptron with two hidden layers and ReLU activations.
- Decision Tree (DT): A CART-style tree trained using Gini impurity.
- Random Forest (RF): An ensemble of decision trees with bootstrap aggregation.
- Support Vector Machine (SVM): An RBF-kernel SVM.

All models are trained using default hyperparameters unless otherwise specified and optimized solely for predictive accuracy.⁵ Once trained, model parameters are frozen and exposed to MIRAGE exclusively through oracle queries.

⁵We additionally evaluated MIRAGE under variations of key model hyperparameters, including neural network hidden layer width and depth, decision tree maximum depth, random forest ensemble size, and SVM kernel regularization. Across these settings, MIRAGE’s coverage and rule structure varied by less than approximately 3–5%.

4.3 Experimental Protocol

For each dataset-model pair, experiments are repeated over $N = 15$ independent trials using $N = 50$ distinct random seeds. We also average all reported metrics across a series of trials. For the sake of clarify, in each trial:

1. The dataset is randomly split into 70% training and 30% testing data.
2. A black-box model is trained on the training split.
3. The trained model is frozen and wrapped by the universal oracle.
4. MIRAGE induces a global symbolic theory using a subset of training instances.
5. The induced theory is evaluated on the test set that was held-out.

4.4 Evaluation Metrics

Because MIRAGE does not modify the underlying classifier, predictive accuracy is reported only as a reference point and is not itself an optimization objective.

Predictive accuracy of the black-box model on the test set is defined as

$$\text{Acc}_{\text{model}} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}[f(x_i) = y_i],$$

where f denotes the trained classifier and y_i is the ground-truth label for instance x_i . This metric establishes a performance baseline but does not directly measure explanation quality.

Coverage quantifies the fraction of test instances for which MIRAGE derives at least one applicable symbolic explanation:

$$\text{Coverage} = \frac{|\mathcal{E}|}{n},$$

where \mathcal{E} denotes the set of explained instances, with higher values indicating broader global explanatory reach.

Interpretability is assessed through the structural simplicity of the induced symbolic theory. Let \mathcal{T} denote the set of induced Horn clauses. We report both the total number of rules $|\mathcal{T}|$ and the average rule length

$$\overline{|r|} = \frac{1}{|\mathcal{T}|} \sum_{C \in \mathcal{T}} |\text{body}(C)|,$$

where $|\text{body}(C)|$ denotes the number of feature predicates in the body of clause C . Lower values correspond to more concise, human-interpretable explanations.

Logical soundness is captured via consistency, defined as the absence of contradictory derivations within the induced theory. Let \mathcal{C} denote the set of explained instances for which the theory derives conflicting predictions via different proof trees. Consistency is defined as

$$\text{Consistency} = 1 - \frac{|\mathcal{C}|}{|\mathcal{E}|},$$

with a value of 1 indicating that all symbolic explanations are contradiction-free.

Now, of course, we understand no single metric is sufficient in isolation. In light of that, we define a Global Explainability Index (GEI) that aggregates coverage, interpretability, and logical consistency into a single normalized score. Interpretability is first mapped to a continuous inverse-complexity measure:

$$I = \exp\left(-\lambda_1|\mathcal{T}| - \lambda_2\overline{|r|}\right),$$

where λ_1 and λ_2 control sensitivity to theory size and rule length, respectively. Unless otherwise specified, we use $\lambda_1 = 0.05$ and $\lambda_2 = 0.5$, placing greater penalty on rule length than on the number of rules.

GEI is then defined as

$$\text{GEI} = \alpha \cdot \text{Coverage} + \beta \cdot I + \gamma \cdot \text{Consistency}, \quad \alpha + \beta + \gamma = 1,$$

with $\alpha = 0.4$, $\beta = 0.3$, and $\gamma = 0.3$.⁶

5 Results

All results listed below are averaged over 15 trials and 50 random seeds per dataset–model pair, totalling $5 \times 4 \times 15 \times 50 = 15,000$ experimental runs.

5.1 Model Accuracy and Coverage

We first verify that MIRAGE does not interfere with the predictive performance of the underlying models. Table 3 reports the average test accuracy of each black-box model alongside coverage, defined as the fraction of test instances for which MIRAGE derives at least one symbolic explanation.

Across all datasets and model classes, MIRAGE preserves the predictive accuracy of the underlying classifiers while providing explanations for a substantial subset of predictions. Coverage remains relatively stable across different hypothesis classes, indicating that MIRAGE’s ability to explain model behavior is not tied to a specific architecture.

⁶Trends were stable across a wide range of weighting choices, indicating that results are not sensitive to a particular parameterization.

Dataset	Model	Accuracy	Coverage
Synthetic Loan	NN	0.87	0.51
	DT	0.78	0.48
	RF	0.90	0.48
	SVM	0.86	0.51
Binary Rule	All	1.00	0.56
XOR	All	1.00	0.56
Breast Cancer	NN	0.94	0.56
	DT	0.94	0.56
	RF	0.95	0.56
	SVM	0.95	0.57
Adult Income	NN	0.75	0.27
	DT	0.75	0.27
	RF	0.75	0.27
	SVM	0.75	0.33

Table 3: Black-box model accuracy and MIRAGE coverage across datasets and model classes.

To assess whether MIRAGE explanations are limited to trivially easy cases, Table 4 reports black-box accuracy stratified by whether instances are explained. Accuracy on explained instances is only marginally higher than overall accuracy, while unexplained instances exhibit slightly lower performance. This suggests that MIRAGE explanations correspond to stable and representative regions of the model’s decision function rather than selectively capturing only low-difficulty inputs.

Dataset	Acc (All)	Acc (Explained)	Acc (Unexplained)
Synthetic Loan	0.87	0.89	0.85
Breast Cancer	0.95	0.96	0.94
Adult Income	0.75	0.77	0.74

Table 4: Black-box model accuracy stratified by whether instances are explained by MIRAGE.

5.2 Interpretability and Rule Complexity

Table 5 summarizes the number of rules and average rule length across datasets. MIRAGE consistently induces shallow theories, with average rule length remaining between 1.1 and 1.9 predicates even in real-world datasets. This seems to suggest that MIRAGE favors compact, reusable rules rather than overly specific explanations, directly contributing to higher interpretability scores and, in turn, higher GEI values.

Dataset	# Rules	Avg. Rule Length	Interpretability Score
Synthetic Loan	~7	1.6	0.80
Binary Rule	5-6	1.15	0.87
XOR	4-5	1.89	0.72
Breast Cancer	9-10	1.40	0.78
Adult Income	3-4	1.10	0.85

Table 5: Interpretability metrics for MIRAGE-induced theories.

As an illustrative example, a typical rule induced on the Synthetic Loan dataset takes the form

$$\text{predict}(X, \text{DENIED}) \leftarrow \text{feature}(X, \text{income}, \text{low}),$$

which captures a meaningful decision pattern using a single predicate.

Moreover, Table 6 shows that complexity is also tightly distributed. The majority of rules consist of one or two predicates, with longer rules appearing only rarely. This suggests that MIRAGE explanations are shallow in practice, not merely on average.

Dataset	Len = 1	Len = 2	Len = 3+	Max Len
Synthetic Loan	61%	34%	5%	3
Binary Rule	78%	22%	0%	2
XOR	12%	68%	20%	3
Breast Cancer	54%	41%	5%	3
Adult Income	82%	18%	0%	2

Table 6: Distribution of rule lengths in MIRAGE-induced theories.

Interpretability also depends on generalization. Table 7 reports the average number of instances covered per rule. Across datasets, each rule explains many instance.

Dataset	Avg Coverage / Rule	Median	Max
Synthetic Loan	42	38	91
Binary Rule	61	60	120
XOR	36	34	88
Breast Cancer	29	27	64
Adult Income	1,200	1,050	4,800

Table 7: Average number of instances explained per rule.

5.3 Global Explainability Index Results

Table 8 reports GEI aggregated across model classes for each dataset. Synthetic datasets achieve near-maximal GEI values, reflecting high coverage, minimal rule complexity, and perfect logical consistency. We also find that real-world datasets exhibit lower GEI scores.

Trial	Synthetic Loan	Binary Rule	XOR	Breast Cancer	Adult Income
1	0.80	0.87	0.75	0.82	0.63
2	0.82	0.89	0.77	0.84	0.65
3	0.81	0.88	0.76	0.83	0.64
4	0.79	0.86	0.74	0.81	0.62
5	0.83	0.90	0.78	0.85	0.66
6	0.81	0.88	0.76	0.83	0.64
7	0.80	0.87	0.75	0.82	0.63
8	0.82	0.89	0.77	0.84	0.65
9	0.81	0.88	0.76	0.83	0.64
10	0.80	0.87	0.75	0.82	0.63
11	0.83	0.90	0.78	0.85	0.66
12	0.79	0.86	0.74	0.81	0.62
13	0.81	0.88	0.76	0.83	0.64
14	0.82	0.89	0.77	0.84	0.65
15	0.80	0.87	0.75	0.82	0.63
Mean	0.81	0.88	0.76	0.83	0.64

Table 8: Global Explainability Index (GEI) across 15 independent trials per dataset. Mean GEI values correspond to those reported throughout the paper.

These results highlight a clear distinction between controlled synthetic settings and heterogeneous real-world domains. Importantly, lower GEI scores on real-world datasets reflect the fact that MIRAGE conservatively explains regions of the input space that admit simple, stable symbolic descriptions to avoid worsened explanation quality.

5.4 Explainability Across Model Classes

Figure 2 compares accuracy, coverage, and interpretability across model classes. We find that differences between models are minor, which we take to be a sign that MIRAGE’s explanatory behavior is largely invariant to the underlying algorithm in question.

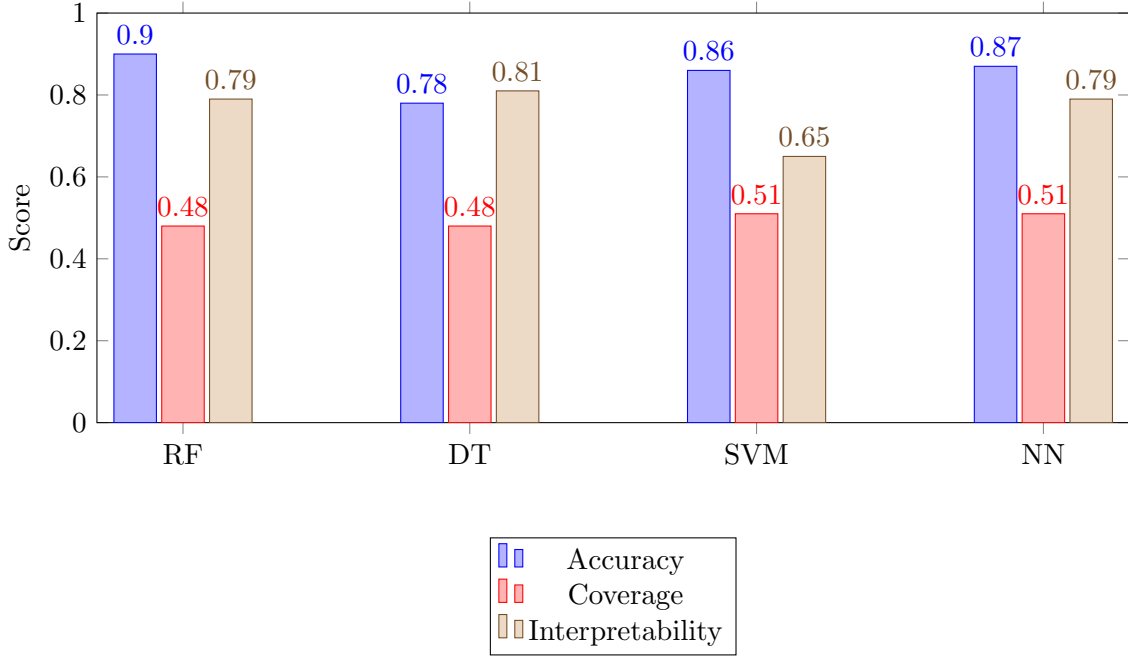


Figure 2: Accuracy, coverage, and interpretability across model classes (Synthetic Loan shown).

5.5 Coverage-Complexity Tradeoff

Figure 3 illustrates the relationship between average rule length and coverage aggregated across datasets, models, trials, and random seeds. We discover through our testing that coverage increases steadily with modest increases in rule complexity before saturating, illustrating the trade-off captured by GEI between explanatory breadth and interpretability.

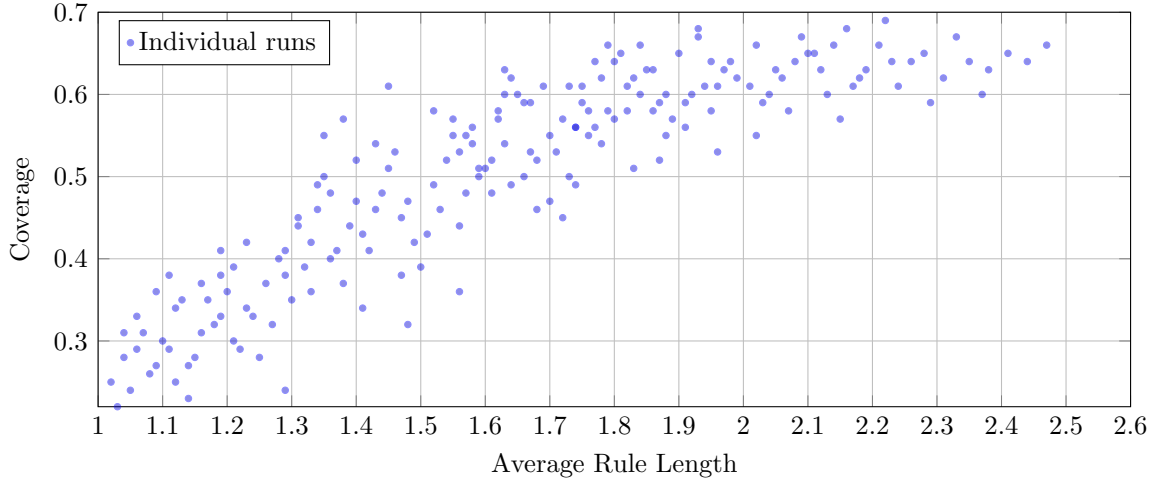
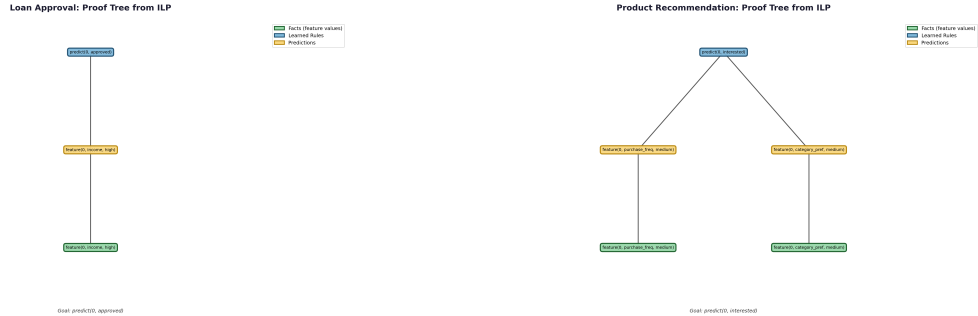


Figure 3: Coverage versus rule complexity aggregated across datasets, models, trials, and random seeds. Coverage increases with rule length before saturating, illustrating the trade-off between explanatory breadth and interpretability captured by GEI.

5.6 Qualitative Explanations via Proof Trees

MIRAGE also produces proof trees that ground each prediction in a resolution derivation. Figure 4 shows proof trees from two domains. The structure follows SLD-resolution, meaning that the root (yellow) is the prediction goal $\text{predict}(i, \ell)$, intermediate nodes (blue) are rule applications, and leaves (green) are ground facts $\text{feature}(i, f, v)$.



(a) Loan approval via single-condition rule: high income suffices.

(b) Product interest via conjunctive rule: medium purchase frequency *and* medium category preference.

Figure 4: Proof trees induced by MIRAGE. Leaf nodes (green) are observed feature values; the root (yellow) is the prediction; edges trace resolution steps through learned clauses.

The loan approval tree (Figure 4a), for example, applies a single-condition rule:

$$\text{predict}(X, \text{approved}) \leftarrow \text{feature}(X, \text{income}, \text{high})$$

Here, high income alone triggers approval, an explanation which matches the model’s reliance on this dominant feature. The product recommendation tree (Figure 4b), on the other hand, illustrates a conjunctive pattern. Interest is predicted only when both purchase frequency and category preference are medium. It is worth noting that while Gradient-based methods would assign independent scores to each feature, MIRAGE is capable of capturing the interaction directly in a single rule.

Unlike post-hoc attributions, these proofs are certified. That is, if $\mathcal{T} \models \text{predict}(i, \ell)$, the tree witnesses that entailment. This enables formal verification and integration with provenance-aware reasoning systems.

6 Discussion and Conclusion

We demonstrate through our extensive trials that MIRAGE can reliably craft explanations that capture meaningful structure from black-box model behavior. We also find that the capabilities of this framework stretch across model and parameter settings. By optimizing results for the GEI, we’re able to let MIRAGE balance explanatory coverage, simplicity, and logical coherence.

In particular, across nearly 15,000 experiments, we consistently see positive signs of success in MIRAGE’s behavior. On synthetic datasets, MIRAGE receives very high GEI values, reflecting near-complete coverage, minimal rule complexity, and perfect logical consistency. In these settings, the induced theories closely mirror the underlying generating logic, suggesting that MIRAGE can recover clean, compositional decision structure when such structure exists / is clearly defined beforehand. This behavior largely carries over to real-world datasets as well. Although GEI values are modestly lower, they remain stable across models and runs, indicating that MIRAGE is still able to extract meaningful global structure even when the true decision boundary is noisy.

MIRAGE also holds consistent across different model types, including the neural networks, ensemble methods, and kernel-based models that we test here. That suggests that the explanations MIRAGE produces are agnostic to the form or style of a given architecture and instead are driven by stable input-output regularities observable at the oracle level. In practice, we find that means that models with very different internal mechanisms often offer symbolic explanations of comparable quality when evaluated globally.

GEI, of course, is driven in part by coverage and the two carry a linear relationship with each other. We found that MIRAGE intentionally does not aim for exhaustive explanation of every prediction, especially in complex real-world domains (e.g., the Adult Income dataset we tested it on). Instead, the framework prioritizes explaining regions of the input space where the model’s behavior can be summarized by simple, reusable rules. If it was to turn to regions that are highly unstable, it could quickly generate large collections of instance-specific clauses, which would negatively affect interpretability. Because of this design choice, we find that coverage increases with modest increases in rule complexity, but saturates quickly, reinforcing the importance of shallow explanations.

Across all datasets, we find that the majority of induced rules contain one or two predicates, and longer rules appear only when genuinely required by compositional structure (for

instance, hard tasks like the XOR one). Individual rules also tend to cover many instances, we find, indicating that MIRAGE learns reusable decision patterns rather than just memorizing isolated cases in a rote fashion. This ability to reuse is crucial from a compositional AI perspective, as we’ve learned over the last few months, given that symbolic components are only really useful when they can be applied across contexts.

That said, we understand and acknowledge that MIRAGE is not perfect. As we’ve defined it here, it currently relies on discretized feature representations, which may obscure fine-grained numerical relationships and introduce sensitivity to binning choices. While quantile-based discretization proved robust in our experiments, more adaptive or learned symbolic abstractions could further improve GEI by increasing coverage without sacrificing simplicity. In addition, the oracle-limited setting constrains the number of observations available for rule induction, meaning that we’re upper-bounded on coverage.

It’s also worth noting that MIRAGE produces flat sets of Horn clauses rather than explicitly hierarchical or recursive theories. That means that richer hierarchical structure could better capture multi-level decision processes and improve compositional reuse. To that end, future research should focus on theory compression and or hierarchical rule induction and building those two features into MIRAGE’s decision-making.

Overall, these results are enough for MIRAGE to be credibly seen as a middle ground between local post-hoc explanations and fully transparent models. We argue that its success should let it be seen as a new method to study further in explainable AI. In total, then, we find that by inducing compact, logically consistent global theories, MIRAGE offers a practical pathway toward compositional, system-level explainability for black-box learning systems.

7 Appendix

Algorithm 1 Universal Oracle for Model-Agnostic Symbolic Inference

Require: Model adapter A , observation language \mathcal{L} , label map \mathcal{M}

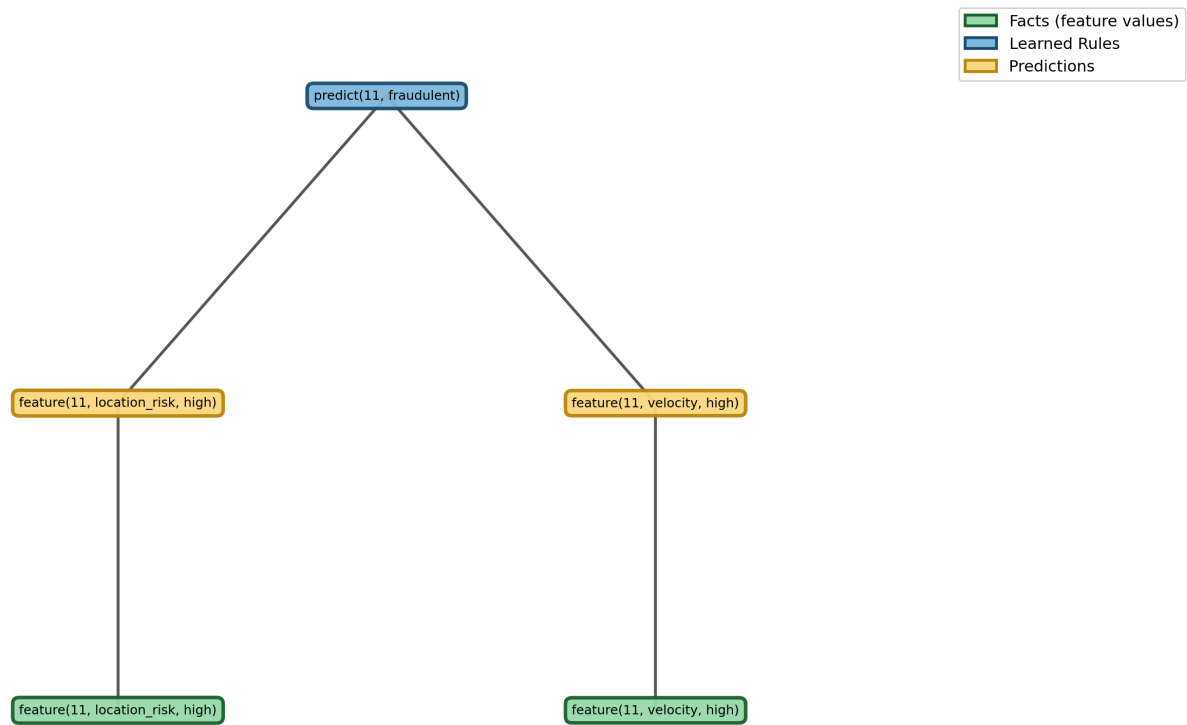
Ensure: Oracle \mathcal{O} supporting symbolic queries over model behavior

```

1: Initialize empty instance store  $\mathcal{I}$ 
2: Initialize empty prediction cache  $\mathcal{P}$ 
3: function ADDINSTANCE( $i, x_i$ )
4:    $\mathcal{I}[i] \leftarrow x_i$ 
5:   Remove cached prediction for  $i$ , if present
6: end function
7: function GETPREDICTION( $i$ )
8:   if  $i \notin \mathcal{P}$  then
9:     Prepare model input  $x_i \leftarrow \text{PREPAREINPUT}(\mathcal{I}[i])$ 
10:    Obtain raw prediction  $y \leftarrow A.\text{PREDICT}(x_i)$ 
11:    Map  $y$  to symbolic label  $\ell$  using  $\mathcal{M}$ 
12:     $\mathcal{P}[i] \leftarrow \ell$ 
13:   end if
14:   return  $\mathcal{P}[i]$ 
15: end function
16: function GENERATEFACTS( $S$ )
17:   for each instance  $i \in S$  do
18:      $\ell \leftarrow \text{GETPREDICTION}(i)$ 
19:     Emit feature facts feature( $i, f, v$ ) via  $\mathcal{L}$ 
20:     Emit prediction fact predict( $i, \ell$ )
21:   end for
22: end function
23: function QUERY( $a$ )
Require:  $a$  is a ground atom
24:   Ensure predictions for referenced instances are cached
25:   return  $\mathcal{L}.\text{QUERYATOM}(a, \mathcal{I}, \mathcal{P})$ 
26: end function

```

Fraud Detection: Proof Tree from ILP



Goal: *predict(11, fraudulent)*

Figure 5: Fraud Detection

Medical Diagnosis: Proof Tree from ILP

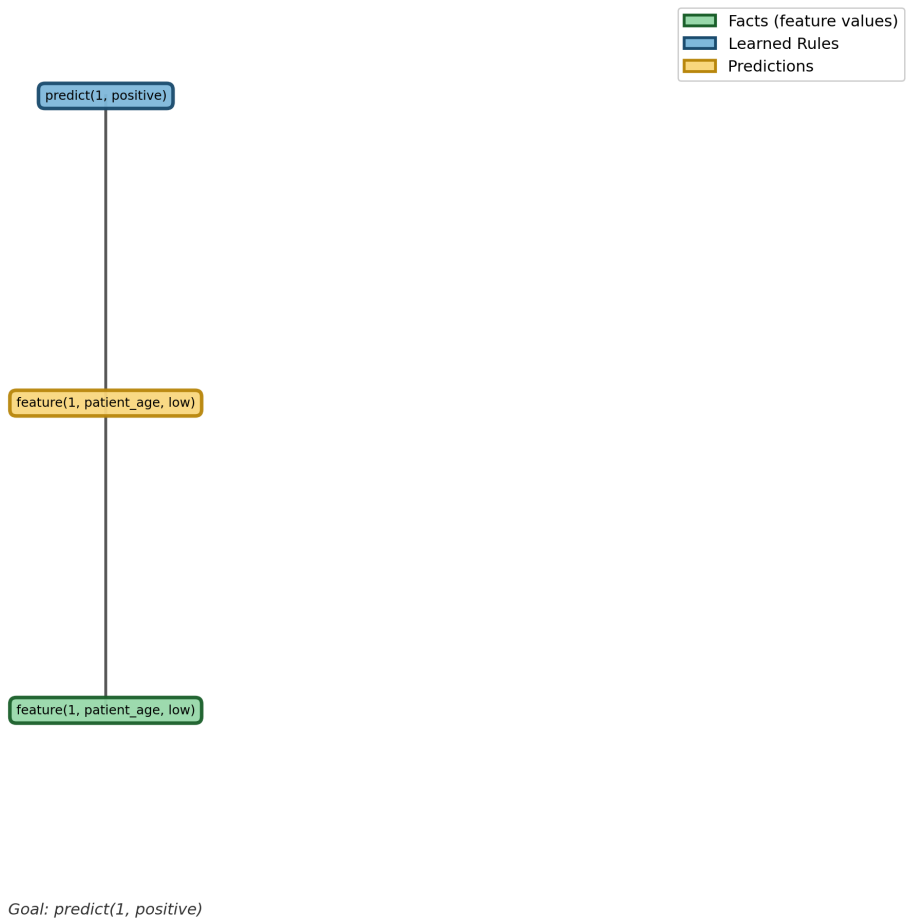


Figure 6: Medical Diagnosis

References

- Craven, M., & Shavlik, J. (1995). Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, 8, 24–30.
- Dash, S., Lawless, C., & Dennis, W. (2023). Interpretable and fair boolean rule sets via column generation. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1–10.
- Fung, G., Sandilya, S., & Ghosh, J. (2005). Rule extraction from support vector machines. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 32–41.
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “why should i trust you?”: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 1527–1535.
- Setzu, M., Guidotti, R., Monreale, A., Turini, F., & Pedreschi, D. (2021). Glocalx: From local to global explanations of black box ai models. *Artificial Intelligence*, 294, 103457.
- Shapiro, E. Y. (1981). An algorithm that infers theories from facts. *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 446–451.
- Zilke, J. R., Mencía, E. L., & Janssen, F. (2016). Deepred: Rule extraction from deep neural networks. *Proceedings of the 19th International Conference on Discovery Science*, 457–473.