



XAI-Units: Benchmarking Explainability Methods with Unit Tests

Jun Rui Lee
Imperial College London
London, United Kingdom
junruilee@yahoo.com

Sadegh Emami
Imperial College London
London, United Kingdom
sadegeh@emami.net

Michael David Hollins
Imperial College London
London, United Kingdom
michaelhollins@hotmail.co.uk

Timothy C. H. Wong
Imperial College London
London, United Kingdom
timchwong1919@gmail.com

Carlos Ignacio Villalobos
Sánchez
Imperial College London
London, United Kingdom
carlosignaciiovillalobos99@gmail.com

Francesca Toni
Imperial College London
London, United Kingdom
ft@imperial.ac.uk

Dekai Zhang
Imperial College London
London, United Kingdom
dz819@imperial.ac.uk

Adam Dejl
Imperial College London
London, United Kingdom
adam.dejl18@imperial.ac.uk

Abstract

Feature attribution (FA) methods are widely used in explainable AI (XAI) to help users understand how the inputs of a machine learning model contribute to its outputs. However, different FA models often provide disagreeing importance scores for the same model. In the absence of ground truth or in-depth knowledge about the inner workings of the model, it is often difficult to meaningfully determine which of the different FA methods produce more suitable explanations in different contexts. As a step towards addressing this issue, we introduce the open-source XAI-UNITS benchmark, specifically designed to evaluate FA methods against diverse types of model behaviours, such as feature interactions, cancellations, and discontinuous outputs.¹ Our benchmark provides a set of paired datasets and models with known internal mechanisms, establishing clear expectations for desirable attribution scores. Accompanied by a suite of built-in evaluation metrics, XAI-UNITS streamlines systematic experimentation and reveals how FA methods perform against distinct, atomic kinds of model reasoning, similar to unit tests in software engineering. Crucially, by using procedurally generated models tied to synthetic datasets, we pave the way towards an objective and reliable comparison of FA methods.

CCS Concepts

• **General and reference** → **Evaluation**; • **Computing methodologies** → *Machine learning*; *Neural networks*; • **Human-centered computing** → *Human computer interaction (HCI)*.

¹The benchmark package is available at <https://github.com/XAI-Units/xaiunits>



This work is licensed under a Creative Commons Attribution 4.0 International License. *FAccT '25, Athens, Greece*
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1482-5/25/06
<https://doi.org/10.1145/3715275.3732186>

Keywords

explainable AI, feature attribution, neural networks, synthetic data, synthetic models, unit testing

ACM Reference Format:

Jun Rui Lee, Sadegh Emami, Michael David Hollins, Timothy C. H. Wong, Carlos Ignacio Villalobos Sánchez, Francesca Toni, Dekai Zhang, and Adam Dejl. 2025. XAI-Units: Benchmarking Explainability Methods with Unit Tests. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*, June 23–26, 2025, Athens, Greece. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3715275.3732186>

1 Introduction

As artificial intelligence (AI) and machine learning (ML) techniques are increasingly embraced, the importance of interpreting these models through explainable AI (XAI) techniques also grows. By improving users' understanding of the logic behind AI models, XAI offers benefits in various settings including increasing social acceptance and trust, meeting legal obligations, detecting and removing bias, debugging unanticipated behaviour and enhancing AI safety.

Feature attribution (FA) methods are a branch of XAI focused on quantifying the effects of input features on model outputs. Common methods include perturbation-based ones such as LIME [33] and SHAP [28], or gradient-based ones such as DeepLIFT [37] and Integrated Gradients [40]. These approaches reduce the complexity of a model's mathematical logic into a set of numerical scores which quantify the importance of each feature.

However, as the number of proposed FA methods has increased, practitioners have encountered confusing situations where these methods contradict each other [34], also known as the *disagreement problem* [23]. This evidently undermines the motivation behind FA methods, which is to disambiguate the reasoning process of an ML model. In response, a variety of metrics have been proposed for evaluating FA methods [30, 48]. However, due to the difficulty of establishing what constitutes a “better” explanation in various scenarios, these metrics are often merely heuristic and may not accurately rank the performance of FA methods.

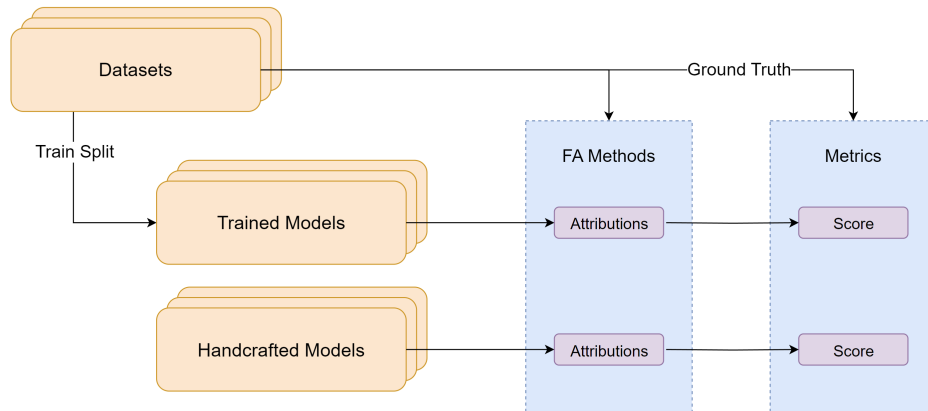


Figure 1: Overview of the XAI-UNITS benchmark. The benchmark provides a set of datasets and models with controlled mechanisms and behaviour. This enables us to evaluate the attributions produced by various FA methods using various metrics, often taking into account the ground-truth expectations associated with the given dataset and model.

With the proliferation of different FA methods and evaluation metrics, a practical need arose to simplify the burgeoning complexity of XAI analysis. Therefore, various XAI toolkits were developed to streamline the comparison between datasets, models, FA methods and metrics [1, 19, 24, 27]. While this has helped the research process, what remains elusive is establishing the conditions under which FA methods reliably capture the internal “reasoning” process of models.

This challenge motivates our XAI-UNITS package (Figure 1), which allows the user to assess how various FA methods perform against expected, atomic units of model behaviour. This highlights the respective strength and limitations of respective FA methods, contributing to greater transparency for users seeking to understand and trust model explanations. By using deterministic models fully aligned with our synthetic datasets and avoiding complex datasets with unclear mechanisms, we are able to provide ground truths facilitating better understanding and evaluation of FA methods. Therefore, while XAI-UNITS focuses on synthetic test cases, this choice is deliberate, enabling us to systematically evaluate FA methods against predictable and distinctive units of model behaviour, which is challenging to achieve with real-world datasets. Although synthetic data simplifies complex real-world scenarios, many of the situations we test, such as feature interaction effects, mirror challenges observed in domains such as healthcare and finance. For those seeking benchmarks on real-world datasets, we refer to complementary efforts in prior work outlined towards the end of Section 2. Moreover, we envision future extensions of our framework to include semi-synthetic or real-world applications.

Our aim with this paper is not to provide definitive reasons as to why some FA methods perform badly in certain settings. Rather, we provide a package which enables researchers to easily find under what conditions particular FA methods struggle, with the hope that this will prompt further exploration into the reasons behind it. To summarise our contributions:

- (1) We provide a benchmark for evaluating FA methods, enabling developers to verify that XAI techniques meet their design specifications and ensuring accountability for correct implementation. Our key approach is to either procedurally generate (“handcraft”) or engineer a collection of neural network models to replicate specific types of testable behaviours.
- (2) We create corresponding synthetic data generators that are paired with each of our models to enable a controlled environment for evaluation.
- (3) We implement the entire benchmarking pipeline within the open source Python library XAI-UNITS, which is fully extensible to support custom evaluation metrics and FA methods.
- (4) We apply our benchmark to common FA methods, testing their strengths and weaknesses on specific model behaviour. Using this approach, we identify an implementation discrepancy in a popular FA library.

The rest of this paper is organised as follows. In Section 2, we describe the related work on evaluating and benchmarking feature attribution methods. Section 3 presents the XAI-UNITS benchmark, and the datasets and models included in it. In Section 4, we report results from applying XAI-UNITS to common FA methods. Finally, in Section 5, we conclude with an overall discussion of the work.

2 Related work

2.1 Evaluation of feature attribution methods

Feature attributions (FA) are a category of XAI methods that calculate attribution scores for all input features for a given model [49]. These scores help to delineate each feature’s impact and importance on the model outcome. Numerous FA methods have been introduced in the literature [6, 13, 28, 32, 33, 37], and can be broadly grouped into two main categories depending on whether they are based on gradients or perturbations [4] (see Speith [39] for a more

Table 1: Comparison between XAI-UNITS and existing toolkits.

Toolkit	Real-World Datasets	Synthetic Datasets	Ground-Truth Available	Extensible	Model Behaviour-focused
XAI-UNITS	No	Yes	Yes	Yes	Yes
OPENXAI [1]	Yes	Yes	No	Yes	No
QUANTUS [19]	No	No	No	Yes	No
$\mathcal{M}^4$ [25]	Yes	Yes	Partial	Yes	No
GRAPHXAI [2]	Yes	Yes	Yes	No	No
XAI-BENCH [27]	No	Yes	Yes	No	No
BAM [44]	No	Yes	No	No	No

detailed review of XAI method taxonomies). Yet the differing approaches may lead to different attribution scores, which is commonly referred to as a *disagreement problem* [23]. This issue has been studied in detail for two widely used explainability methods, LIME and SHAP [34]. Thus, to assess the quality of the attribution scores across FA methods, there is a clear need for their reliable evaluation.

Several evaluation metrics for FA methods have been proposed following questions around the effectiveness and consistency across different FA methods [7, 23]. By evaluating FA methods, researchers and users may get a better idea as to which method performs better in a particular case. Accordingly, different evaluation metrics are used to test FA methods against certain desirable properties [28, 40]. For instance, for a given FA method, the *explanation infidelity* metric aims to capture its faithfulness, whilst the *explanation sensitivity* (or *max-sensitivity*) metric measures its robustness [3, 45]. That said, these metrics typically only offer a rough indication of an FA method’s performance relative to other competing methods: in most cases, there is no unambiguous metric indicating the absolute quality of any FA method due to the absence of ground truth attributions. Some FA evaluation methods are specific to certain modalities, such as the metrics based on 2x2 image grids proposed by Fresz et al. [16] for evaluating explanations in computer vision.

2.2 Synthetic approaches

Synthetic datasets are utilised in many of the existing benchmarks for FA methods because ground truth attributions can be derived from them [2, 5, 47, 49]. Different benchmarks generate their synthetic datasets differently. XAI-Bench constructs their tabular synthetic datasets via mimicking common statistical distributions [27] and the ground truth attributions are derivable from the corresponding distribution. Similarly, the synthetic data generator from OPENXAI is based on sampling clusters from Gaussian distributions, which, given the most accurate possible model, has been proven to facilitate the computation of ground-truth attributions for each cluster [1]. However, there is no guarantee that these “ground-truth” attributions will be aligned with imperfect models that result from the used training procedures. The tabular benchmarks in XAI-UNITS are also based on synthetic data sampled from statistical distributions, but in contrast to the packages mentioned above, each dataset is paired with a handcrafted model with perfect accuracy. Additionally, XAI-UNITS datasets and models are focused on atomic model behaviours rather than generic statistical distributions.

Handcrafted neural network models are rare in most existing benchmarks because in real-world applications, models are trained to fit the observed data rather than predefined. However, as our goal is not to replicate real-world scenarios but to create a controlled environment to test FA methods, we pair synthetic datasets with handcrafted models. As argued by Brieman [47], this is motivated by the *Rashomon effect* [8], where the ground-truth attributions derived from synthetic datasets are trustworthy and effective only when they are paired with handcrafted models during the evaluation of FA methods. With our focus on using ground-truth attributions to evaluate FA methods, our approach is related to the *Synthetic Explainable Classifier* generators of Guidotti [18]. However, our focus is on how FA methods perform against particular, controlled, distinct kinds of model reasoning. Therefore, our approach offers several potential benefits to the XAI community, such as allowing researchers to test how well new FA methods handle particular types of feature interaction.

2.3 Toolkits for evaluating feature attribution methods

Toolkits for FA methods have been developed for the easy application of evaluation metrics and benchmarking across different FA methods. The benchmarks offered by existing toolkits are generally constructed based on either real-world [12, 20, 26, 46] or synthetic datasets [27, 29, 44], or a blend of both [1, 2, 25]. According to a recent survey on existing toolkits, OPENXAI and QUANTUS are two of the most popular options [24]. OPENXAI [1] constructs its benchmark by applying FA methods to trained models on real-world datasets. A synthetic data generator that generates multiple clusters of normally-distributed data is also available for benchmarking. Moreover, it offers an open-source end-to-end Pipeline for implementing FA methods and evaluating them. Meanwhile, QUANTUS [19] is a Python package that gathers a diverse pool of over 30 different built-in evaluation metrics and is extendable to custom evaluation metrics. Although QUANTUS does not contain any pre-loaded datasets, the user can load in their own data and use the implemented metrics to evaluate FA methods and other XAI methods with respect to various properties.

Sharing commonalities with OPENXAI and QUANTUS, XAI-UNITS provides a complete Pipeline for benchmarking FA methods and is able to support custom methods and metrics during evaluation. However, our work distinguishes itself from the existing toolkits

Table 2: Summary of available synthetic data generators in XAI-UNITS.

Data Generator	Datatype	Feature type(s)	Ground Truth	Default Metric
Weighted Continuous	Tabular	Continuous	Available	MSE
Conflicting Features	Tabular	Continuous, Categorical	Available	MSE
Pertinent Negatives	Tabular	Continuous, Categorical	Available	MSE
Feature Interaction	Tabular	Continuous, Categorical	Available	MSE
Uncertainty	Tabular	Continuous	Mask	Mask Error
Shattered Gradient	Tabular	Continuous	Unavailable	SensitivityMax
Boolean Formula	Tabular	Categorical	Unavailable	Infidelity
Boolean AND	Tabular	Categorical	Available	MSE
Boolean OR	Tabular	Categorical	Available	MSE
Balanced Image	Images		Mask	Mask Proportion Image
Imbalanced Image	Images		Mask	Mask Proportion Image
Trigger Injection	Text		Mask	Mask Proportion Text

by enabling the evaluation of FA methods in a more controlled setting. Enabled by the usage of procedurally generated handcrafted models with known internal mechanisms, each dataset and model pair is analogous to a unit test that isolates a single type of input behaviour to test for. This effectively circumvents the *blame problem* [31], as our toolkit eliminates the ambiguity in deciding whether a FA method’s poor performance is driven by the method itself or the model behaviour. Moreover, XAI-UNITS is compatible with a broad range of data modalities and model architectures. In particular, the toolkit incorporates multilayer perceptrons (MLPs), convolutional neural networks (CNNs), vision transformers (ViTs) and large language models (LLMs), while also supporting diverse modalities including tabular data, images and text.

3 Package overview

Here, we describe XAI-UNITS’ core components, as outlined in Figure 1. We focus on the datasets and models included in the benchmark, followed by the tested FA methods and the used metrics.

3.1 Datasets and models

XAI-UNITS contains seven tabular, two image, and one text synthetic data generators, summarised in Table 2. Each tabular dataset generator is paired with a handcrafted neural network model whose logic is set out formally in Appendix A. Motivating each of these pairs are distinct, simple units of behaviour which some FA methods may struggle with. Clearly, this set of pairs is not exhaustive, but by making XAI-UNITS easily extensible, we encourage other researchers to add their own “unit tests”. Apart from the handcrafted models, the benchmark also provides analogous trained models for comparison.

Weighted Continuous. This neural network returns a weighted sum of input features. It serves as a simple baseline for evaluating FA methods and a springboard for developing the more complicated models. It implements a linear function using a two-layer MLP network with ReLU activations in the hidden layer.

Conflicting Features. This dataset-model pair introduces categorical “cancellation” features which cancel or negate the impact of continuous features. This tests how well FA methods handle

cases where there are conflicts between features. Aim to surface such conflicts has previously informed the design of several FA methods, including DeepLIFT RevealCancel [38] and CAFE [14]. As an example of a conflict between features, consider a healthcare AI system predicting a patient’s risk of death based on vital signs and previous treatments. When faced with a normal temperature reading, the system may typically predict a lower overall risk, but this line of reasoning may be void when the patient was recently administered an antipyretic drug.

Pertinent Negatives. This dataset-model pair captures scenarios where an output meaningfully depends on the zero value of a (pertinent negative) feature. Such feature behaviour may be present in models that predict the likelihood that a patient is suffering from a severe heart condition given the heart rate. As the 0 heart rate is meaningful for the prediction (indicating asystole), FA methods should ideally return non-zero attributions for this feature. However, this may trouble FA methods as the 0 feature value may lead to a 0 attribution score being returned.

Shattered Gradients. The logic of shattered gradients is that minor input changes that have negligible impact on the model output can lead to significant changes in attribution scores. An example of this is the point of gradient discontinuity in the following function $\text{ReLU}(x - 100)$. Any infinitesimal positive perturbation around the discontinuity will still result in similar output however, gradients and thus attribution would change drastically relative to the magnitude of the perturbation.

Categorical Feature Interaction. This dataset-model pair is based on interactions between categorical and continuous features. Each continuous feature’s weight varies depending on the associated categorical feature’s value — either 0 or 1. Specifically, for any given pair consisting of a continuous feature and a categorical feature, the weights are defined as $(w^{(1)}, w^{(2)})$. If the categorical feature’s value is 0, the weight applied to the continuous feature is $w^{(1)}$. Conversely, if the categorical feature’s value is 1, the weight becomes $w^{(2)}$. An example of the interacting features logic is in predicting a client’s credit score; the importance of their salary may depend on (or interact with) whether they are “old” or “young”.

Uncertainty Model. This model-dataset pair captures when a subset of input features is irrelevant for probabilistic class prediction. In the case of a classifier that is given redundant inputs (with no impact on the prediction), one would expect a perfect FA method to not assign attribution scores to these inputs.

The model is composed of a linear transformation followed by a softmax activation layer. In this model, some input features are irrelevant to output class prediction and so are designated as common. These common features simply add a constant term to all output class logits equally. Thus, as the softmax layer is translation invariant $f(x + b) = f(x)$ where x, b are vectors, common features have no impact on the output class prediction. Thus the default evaluation metric is Mask Error (defined in Appendix A.7) which penalises explanations that give larger attributions to the common features.

Boolean Formula. The final tabular datasets and models provided are Boolean formulae. Using this framework, the user may implement neural networks that replicate the logic of any arbitrary Boolean formula that is made up of the 'AND', 'OR', and 'NOT' connectives. The dataset for any such formula consists of permutations of truth values for the propositional atoms. Each propositional atom is represented as +1 if its value is True, or as -1 if its value is False. The significance of this model-dataset pair is given by the common existence of sufficient/necessary conditions for a particular prediction and the general utility of logical rules in reasoning.

Image datasets. In addition to the tabular datasets, we also provide image datasets which are designed to overlay various foregrounds (geometric shapes or images of dinosaurs) onto diverse, textured backgrounds.² This setup is crucial for creating scenarios that closely mimic real-world conditions where objects of interest (foregrounds) appear against varying scenes (backgrounds). A good FA method applied to a well-trained model might assign higher attribute scores to pixels in the foreground rather than the irrelevant background.

For a given sample (combining a foreground and background), each image can be customised in terms of the position, size, rotation, and colour of the foreground objects (see Figure 2). These attributes can be set to fixed values or varied randomly, thereby introducing necessary variations that challenge the robustness of image recognition models. In addition, the library has the ability to generate both balanced and imbalanced datasets. In a balanced dataset, every combination of background, foreground, and colour appears an equal number of times, which is ideal for basic model training where equal representation ensures unbiased learning. The imbalanced dataset, conversely, simulates real-world scenarios where certain objects might appear more frequently with specific backgrounds, causing the model to focus more on the background with the attribution scores changing accordingly.

Text datasets. Given the rapid adoption and evolution of LLMs, our benchmark also includes a dedicated text dataset and corresponding LLMs, enabling practitioners to experiment with the application of FA methods on these models. However, using standard

²Dinosaur image files were sourced from Wikimedia Commons, the free media repository, and licensed under the Creative Commons license CC BY-SA [41]. Textured backgrounds were taken from the Describable Textures Dataset (DTD), available to the computer vision community for research purposes [10].

LLM and generic prompts would prohibit us from objectively comparing the attribution scores across the different FA methods, as we are unable to objectively identify which tokens are necessary for the next token generation.

In line with the philosophy of our package, we provide a "unit test" for FA method on LLMs called "Trigger Injection", drawing inspiration from Saha et. al. [35] and Yan et. al. [43]. "Trigger Injection" has two components, a dataset of prompts with Trigger Words embedded within, and a fine-tuned Llama-3.2-1B-Instruct called TriggerLLM.

TriggerLLM has been fine-tuned such that, in the presence of the Trigger Word in a prompt, the LLM will respond by only generating the Trigger Response Token. Otherwise, in the absence of the Trigger Word in a prompt, TriggerLLM will generate responses as per usual. Training the model to respond to the trigger word provides clear expectations of the model's behaviour, thus enabling the direct comparison of attributions scores. We report further details about our fine-tuning process for TriggerLLM in Appendix A.10. Apart from being a useful model of data backdoor attacks, our "Trigger Injection" unit test can also serve as a way to isolate the effects of specific instructions in the prompt on the model behaviour.

3.2 Feature attribution methods and evaluation metrics

Having covered the dataset and model components, here we briefly review the integrated FA methods and metrics. To ensure compatibility with the existing ecosystem, the XAI-UNITS package natively supports running CAPTUM [22] FA methods (e.g. DeepLIFT, ShapleyValueSampling [9]) and also contains a wrapper class for running custom FA methods. The package also supports CAPTUM's official attribution wrappers for LLMs. Similarly, our package supports the evaluation metrics from CAPTUM (Infidelity, SensitivityMax), as well as providing a wrapper class for running custom metrics. Table 2 displays the default evaluation metric used for each dataset in the package: in cases where a ground truth attribution can be determined, mean squared error (or a variant thereof) is used as the default metric. Our ground truth attributions were derived by ablating inputs to a baseline reference value (in most cases 0). The full definitions and the associated details are provided in Appendix A.

4 Benchmarking analysis

We now demonstrate how experiments run with the XAI-UNITS package can provide novel insights for evaluating FA methods. The code and instructions to reproduce all results in this section are provided in the supplementary materials.

4.1 Tabular dataset experiments

To start with, we tested the performance of several common FA methods on the tabular datasets and handcrafted models outlined in Section 3.1. For comparison, we ran identical experiments for both our handcrafted model and trained models.

Experiment setup. The FA methods we experimented with are (the CAPTUM [22] versions of) DeepLIFT [37], InputXGradient [38],

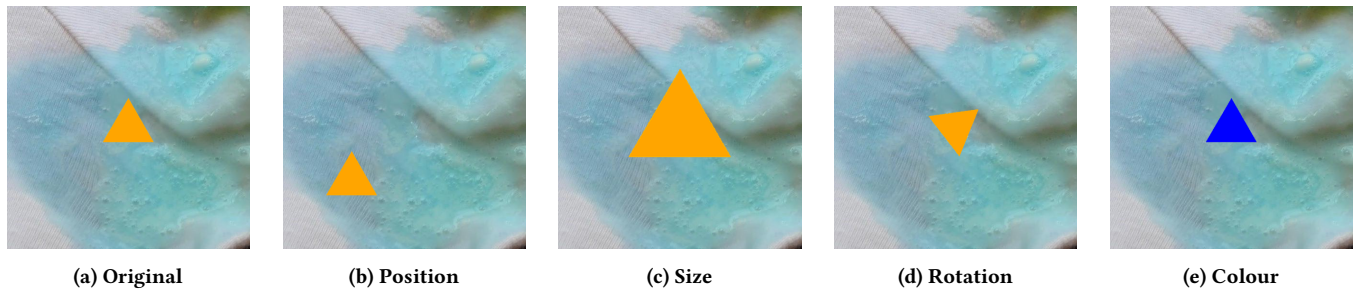


Figure 2: Variations of a foreground-background combination from the BalancedImageDataset.

IntegratedGradients [40], LIME [33] (with linear regression without regularisation as the surrogate model, and with lasso regression), KernelSHAP [28] and ShapleyValueSampling [9]. We initialise the datasets introduced in Section 3 and split them into train, validate and test subsets (with 2600, 400, and 1000 data points respectively). We use the test subset for the FA evaluations. Each dataset has ten input features (except the ConflictingDataset, which has ten additional “cancellation” features) with 1000 data points for evaluation. The trained model was a ReLU MLP with three hidden layers, each 100 neurons wide. FA methods were evaluated using the default metric of each dataset. This experiment was repeated for five trials using different random model initialisations. All models were trained on a single retail GPU (RTX 4070Ti) and experiments were run on a single retail CPU (AMD Ryzen 9 7950X).

High-level findings. Table 3 summarises the results. We have given the mean and standard deviations for FA methods evaluated across the five trials, and the model performance for the trained models (at the bottom of the table). All FA methods performed well on the simplest test case, the Weighted Continuous models, but struggled on models with gradient discontinuities, such as those for Shattered Gradients and Pertinent Negatives. In addition, in line with intuition, FA methods that rely on linear surrogate models such as KernelSHAP and LIME tend to perform worse on significantly non-linear models. Our results also indicate that FA methods generally perform worse on the trained models compared to the handcrafted models (with notable exceptions as discussed in the case studies below). This is expected, as imperfect, trained models may not be fully aligned with the ground truth. Thus, given the higher prediction error of the trained models (see Model Performance score) we would expect a higher attribution error. We also note the poor results for InputXGradients on certain models — this is likely due to gradients only being representative of the local model behaviour and not capturing the full effects of the input features. Following these general observations, some specific results may be of interest.

Case study: Conflicting behaviour. Running the experiments on the Conflicting Feature models, we notice that gradients-based FA methods performed significantly worse on the handcrafted models in comparison with the trained models. We hypothesise this is due to the inherent limitations of gradient-based FA methods, which are exacerbated by our handcrafted model.

Gradient-based FA methods are known to struggle with propagating importance signals when gradients are zero [37]. In our handcrafted Conflicting Feature model implementation (see Section 3 and Appendix A.3 for details), conflicting features push the gradients of the hidden layers to zero, resulting in zero attribution for both features involved in the conflict. This issue, zero-gradients leading to zero attribution scores, is also prevalent in any neural network with ReLU layers, including the trained models used. However, we observe that the effect is less pronounced in trained neural networks. We speculate that this is likely due to the conflict behaviour being distributed across multiple neurons, which would reduce the prevalence of truncated gradient signals during backpropagation. In contrast, perturbation-based methods do not seem to be affected by this problem. ShapleyValueSampling, in particular, is one of the best-performing methods for Conflicting Feature models. This phenomenon is likely attributable to its considerations of various combinations of inputs, which can reveal possible feature conflicts even if the network gradients are zero. However, merely relying on perturbations for computing attributions does not appear to be sufficient for accurately capturing the effects of conflicting features. Notably, FA methods using linear surrogate models also struggle with the Conflicting Feature models due to the non-linearity of the model output.

Case study: Implementation discrepancy. The experiments run on the Uncertainty models highlight a discrepancy between the original design of DeepLIFT and the CAPTUM implementation. The DeepLIFT paper recommends that, in the case of Softmax outputs, we may prefer to compute contributions to the logits rather than contributions to the Softmax outputs. *If* we compute the contributions to the logits, then the paper also recommends applying a normalisation step. However, the CAPTUM implementation applies the normalisation step whenever we compute contributions to the Softmax outputs and does not apply the normalisation step when computing contributions to the logits.

We expect DeepLift to achieve a minimal Mask Error on the UncertaintyAwareDataset but the default CAPTUM implementation gives a high score (Table 4). By using the correct target layer for the attributions (logits) and applying the normalisation step, we see a perfect score. Our analysis demonstrates the utility of XAI-UNITS not only for comparing different FA methods, but also for surfacing and diagnosing issues with their implementation.

Table 3: Tabular Dataset Results. ↓ / ↑ indicates a low / high score is better.

	Weighted Fts ^a	Default Metric [↓]		Uncertainty ^d
		Conflicting ^b	Interacting ^c	
DeepLIFT				
Handcrafted	0.000 ± 0.000	0.175 ± 0.040	0.000 ± 0.000	62.915 ± 80.537
Trained	0.003 ± 0.002	0.061 ± 0.023	0.089 ± 0.076	0.002 ± 0.000
InputXGradient				
Handcrafted	0.000 ± 0.000	0.175 ± 0.040	0.000 ± 0.000	0.000 ± 0.000
Trained	0.006 ± 0.003	0.108 ± 0.026	0.124 ± 0.113	0.005 ± 0.001
IntegratedGradients				
Handcrafted	0.000 ± 0.000	0.175 ± 0.040	0.000 ± 0.000	0.000 ± 0.000
Trained	0.003 ± 0.002	0.060 ± 0.023	0.084 ± 0.070	0.006 ± 0.002
KernelSHAP				
Handcrafted	0.000 ± 0.000	0.470 ± 0.116	0.128 ± 0.108	0.000 ± 0.000
Trained	0.002 ± 0.001	0.574 ± 0.113	0.131 ± 0.105	0.007 ± 0.002
ShapleyValueSampling				
Handcrafted	0.000 ± 0.000	0.055 ± 0.013	0.073 ± 0.067	0.000 ± 0.000
Trained	0.001 ± 0.000	0.068 ± 0.022	0.074 ± 0.067	0.001 ± 0.000
LIME (Linear)				
Handcrafted	0.000 ± 0.000	0.179 ± 0.040	0.099 ± 0.082	0.000 ± 0.000
Trained	0.001 ± 0.001	0.253 ± 0.059	0.104 ± 0.084	0.004 ± 0.001
LIME (Lasso)				
Handcrafted	0.005 ± 0.001	0.092 ± 0.022	0.090 ± 0.075	0.000 ± 0.000
Trained	0.006 ± 0.001	0.113 ± 0.030	0.093 ± 0.075	0.001 ± 0.000
Model Performance ^e				
Handcrafted	0.000 ± 0.000 [↓]	0.000 ± 0.000 [↓]	0.000 ± 0.000 [↓]	1.000 ± 0.000 [↑]
Trained	0.006 ± 0.004 [↓]	0.160 ± 0.081 [↓]	0.032 ± 0.026 [↓]	0.944 ± 0.012 [↑]
	Shattered Grad ^f	Pertinent Neg ^g	Bool AND ^h	Bool OR ⁱ
DeepLIFT				
Handcrafted	1.896 ± 0.147	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Trained	17.510 ± 8.520	0.351 ± 0.520	0.045 ± 0.034	0.033 ± 0.025
InputXGradient				
Handcrafted	1.896 ± 0.147	11.899 ± 4.649	0.094 ± 0.002	3.287 ± 0.005
Trained	96.579 ± 49.496	0.953 ± 1.419	0.066 ± 0.002	0.066 ± 0.004
IntegratedGradients				
Handcrafted	1.896 ± 0.147	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
Trained	17.943 ± 8.438	0.345 ± 0.514	0.044 ± 0.035	0.034 ± 0.025
KernelSHAP				
Handcrafted	2.089 ± 0.364	0.000 ± 0.000	0.355 ± 0.008	0.352 ± 0.004
Trained	4.652 ± 1.020	0.301 ± 0.490	0.187 ± 0.157	0.180 ± 0.153
ShapleyValueSampling				
Handcrafted	0.825 ± 0.027	0.000 ± 0.000	0.010 ± 0.000	0.010 ± 0.000
Trained	0.999 ± 0.253	0.218 ± 0.403	0.045 ± 0.033	0.034 ± 0.024
LIME (Linear)				
Handcrafted	1.972 ± 0.719	0.000 ± 0.000	0.064 ± 0.001	0.064 ± 0.001
Trained	55.196 ± 26.218	0.247 ± 0.424	0.072 ± 0.008	0.063 ± 0.005
LIME (Lasso)				
Handcrafted	2.796 ± 1.055	0.004 ± 0.000	0.058 ± 0.001	0.059 ± 0.001
Trained	2.825 ± 0.877	0.230 ± 0.403	0.076 ± 0.016	0.068 ± 0.011
Model Performance ^j				
Handcrafted	0.000 ± 0.000 [↓]	0.000 ± 0.000 [↓]	0.000 ± 0.000 [↓]	0.000 ± 0.000 [↓]
Trained	0.003 ± 0.003 [↓]	1.100 ± 2.022 [↓]	0.001 ± 0.003 [↓]	0.002 ± 0.003 [↓]

^a WeightedFeaturesDataset (MSE[↓])^b ConflictingDataset (MSE[↓])^c InteractingFeatureDataset (MSE[↓])^d UncertaintyAwareDataset (Mask Error[↓])^e The metric for Model Performance is MSE[↓] except for UncertaintyAwareDataset using Accuracy[↑].^f ShatteredGradientsDataset (SensitivityMax[↓])^g PertinentNegativesDataset (MSE[↓])^h BooleanAndDataset (MSE[↓])ⁱ BooleanOrDataset (MSE[↓])^j The metric for Model Performance is MSE[↓] except for UncertaintyAwareDataset using Accuracy[↑].

Table 4: DeepLIFT on UncertaintyAwareDataset. Varying the target. ↓ / ↑ indicates a low / high score is better.

	Mask Error↓		
	Outputs ^a	Logits	Logits with Normalisation
Handcrafted	62.9 ± 80.5	0.991 ± 0.016	0.000 ± 0.000

^a CAPTUM default implementation includes a normalisation bug.

4.2 Image dataset experiments

Briefly, we also consider results on the image datasets, balanced and unbalanced, summarised by Table 5. The same regime was used for both datasets — generating 3000 images with an 80/10/10 split where the test set is used for evaluating FA methods. A 1 million-parameter CNN and a Vision Transformer (ViT) with about 2 million parameters were used in the experiment, both randomly initialized across 5 seeds. These models were chosen to compare how different architectural designs impact performance and interpretability. CNNs rely on convolutional operations, which introduce inductive biases allowing them to efficiently detect local features across the image. In contrast, ViTs use attention mechanisms to capture global relationships between image patches but lack these inductive biases [42], making them less naturally translation invariant. As might be expected, for the imbalanced dataset, the model learns to rely on the background in order to achieve higher accuracy. As a consequence, we see that the FA methods assign a large proportion of the FA to the background rather than to the foreground shape. A CNN model trained with balanced backgrounds does not experience this problem so more than 90% of the attributions are inside the shape mask. However, a ViT trained on balanced backgrounds shows significantly worse results, with approximately 70% of the attributions falling inside the shape mask. We hypothesize that this might be caused by a lack of inductive bias for the locality in ViT, as the model can freely attend to arbitrary input patches instead of integrating information from progressively larger neighbourhoods as done by a CNN. The results allow us to compare how well different feature attribution methods discriminate between models focused on the true signal and those relying on spurious correlations with the background.

4.3 Text dataset experiments

In the experiment with textual data, we compared five FA Methods, using our text dataset with two versions of the fine-tuned LLM: TriggerLLM and TriggerLLM_{deterministic}. The latter model is optimised to more reliably respond to the trigger token at the expense of a more substantial drop in performance on other tasks — see Appendix A.10 for more information. We focused on evaluating the FA methods applicable to LLMs using the official CAPTUM wrappers. In all cases, the baseline token chosen was set to " " or white-space token rather than the zero vector or zero token id, as we considered this to be a more natural and neutral baseline for textual data.

FeatureAblation and ShapleyValueSampling were the two best-performing FA Methods for both LLMs. However, it is worth noting that ShapleyValueSampling takes an order of magnitude more time to run compared to other FA methods, hence only 100

samples were used for its experiment. The FA Methods based on linear surrogate models struggled due to the non-linear nature of the model/dataset but still produced reasonable attribution scores.

IntegratedGradients was the worst-performing method. As Sanyal et. al. mention [36], the straight-line interpolation, used in IntegratedGradients, may not be appropriate given LLM inputs are discrete units and there are no intermediate states between two tokens, which can lead to inaccurate attributions.

It is also noteworthy that the relative ranking of the FA methods evaluated on the two different LLMs was consistent.

5 Conclusion and discussion

Within the XAI community, there is currently no consensus on the universally best approach for evaluating FA methods. While there are many existing benchmarks for this purpose, the benchmark we developed is unique in its focus on atomic “test cases” and comparison with ground-truth attribution scores. We achieve this by providing pairs of synthetic datasets and handcrafted neural network models. This creates a set of calibrated input interactions and model behaviours, against which we can evaluate FA methods using a battery of evaluation metrics. We do not claim that our benchmark offers a definitive ranking of FA methods. Rather, we argue that it is an effort towards address the *disagreement problem* [23] and provides valuable insights into specific behaviours that can naturally complement results from other benchmarks and evaluation approaches. These insights can inform users about the strengths and weaknesses of FA methods in specific settings, improving transparency in their application. Furthermore, isolating scenarios where a particular FA method may falter can verify whether the design specifications of the method are fulfilled, thereby promoting accountability for the developer. This is exemplified by our case study in Section 4.1, which highlights the discrepancy in the implementation of DEEPLIFT in CAPTUM.

Our benchmark is accessible in the form of the XAI-UNITS Python package, which is fully open-source and extensible to custom feature attribution methods or evaluation metrics. Moreover, XAI-UNITS has been designed such that the evaluation procedure is streamlined and researchers can effortlessly run their experiments on multiple FA methods.

A potential limitation of our work is that the performance of FA methods on handcrafted neural networks is not guaranteed to represent their performance in the real world [31]. Nevertheless, we believe in the merits of our approach. First, using synthetic models is the only way to guarantee model alignment with the data distribution and expected behaviour. Second, our models are directly modelling real-world scenarios, e.g., interacting or conflicting features. Finally, we also provide trained models for comparison, which enables us to see how real networks behave under controlled conditions while loosening the alignment guarantees.

While the experiments conducted in this report showcase that our benchmark can be used to evaluate the correctness of FA methods, we note that other properties of FA methods, such as the *Co-12* properties introduced by Nauta et. al. [30], need to be evaluated to ensure holistic assessment of FA methods. Although our benchmark is designed to be compatible with custom FA methods evaluation

Table 5: Image dataset results. ↓ / ↑ indicates a low / high score is better.

	Mask Proportion Image [↑]		ViT	
	Balanced Images	Imbalanced Images	Balanced Images	Imbalanced Images
DeepLIFT	0.944 ± 0.048	0.522 ± 0.241	0.708 ± 0.024	0.575 ± 0.040
InputXGradient	0.955 ± 0.041	0.497 ± 0.209	0.708 ± 0.024	0.575 ± 0.040
IntegratedGradients	0.949 ± 0.046	0.531 ± 0.215	0.675 ± 0.019	0.491 ± 0.016
Test Accuracy [↑]	0.862 ± 0.120	0.925 ± 0.112	0.818 ± 0.122	0.958 ± 0.089

Table 6: Text dataset results. ↓ / ↑ indicates a low/high score is better.

	Samples	Mask Ratio [↑]	
		TriggerLLM	TriggerLLM _{Deterministic}
FeatureAblation	1000	0.078 ± 0.102	1.000 ± 0.000
IntegratedGradients	1000	-0.006 ± 0.138	0.040 ± 0.239
LIME	1000	0.018 ± 0.056	0.367 ± 0.113
KernelSHAP	1000	0.013 ± 0.077	0.135 ± 0.168
ShapleyValueSampling	100	0.057 ± 0.063	0.669 ± 0.066

metrics (thus supports multi-faceted evaluations), ultimately it is limited to a technical evaluation of FA methods.

Human interpretability is an increasingly important property of XAI methods [21] and many additional factors need to be considered for human-centric evaluations, such as explanation complexity (affecting how understandable they are to humans) and the role of explanations in effective human-AI interaction. Addressing these aspects may require user studies and human-AI performance evaluations, which are out of the scope of our current benchmark. As potential future work, developing a user-friendly graphical interface could enhance the accessibility and usability of our benchmark for a broader range of researchers and practitioners.

Overall, to the best of our knowledge, our work is the first within the research community that provides an end-to-end pipeline to benchmark FA methods via a diverse set of synthetic datasets and handcrafted models. With ease of use and transparency in modelling, we hope that our work will aid researchers and practitioners alike in gaining a better understanding of the strengths and limitations of various FA methods.

Acknowledgments

We thank Theo Reynolds for his involvement and contributions throughout the development of this project.

This research was partially supported by ERC under the EU’s Horizon 2020 research and innovation programme (grant agreement no. 101020934, ADIX), by J.P. Morgan and the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme (grant agreement no. RCSR2021\11\45) and by UKRI through the CDT in AI for Healthcare <https://ai4health.io/> (grant agreement no. EP/S023283/1).

References

- [1] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. OpenXAI:

- Towards a Transparent Evaluation of Model Explanations. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/65398a0eba8c9b4a1c38ae405b125ef-Abstract-Datasets_and_Benchmarks.html
- [2] Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. 2023. Evaluating Explainability for Graph Neural Networks. *Scientific Data* 10, 144 (2023). <https://www.nature.com/articles/s41597-023-01974-x>
- [3] David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the Robustness of Interpretability Methods. *arXiv:1806.08049* <http://arxiv.org/abs/1806.08049>
- [4] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=Sy21R9JAW>
- [5] Leila Arras, Ahmed Osman, and Wojciech Samek. 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion* 81 (2022), 14–40.
- [6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one* 10, 7 (2015), e0130140.
- [7] Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. 2024. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences* 121, 2 (Jan. 2024). doi:10.1073/pnas.2304406120
- [8] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.
- [9] Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research* 36, 5 (2009), 1726–1730.
- [10] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing Textures in the Wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23–28, 2014*. IEEE Computer Society, 3606–3613. doi:10.1109/CVPR.2014.461
- [11] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *CoRR* abs/2110.14168 (2021). [arXiv:2110.14168](https://arxiv.org/abs/2110.14168) <https://arxiv.org/abs/2110.14168>
- [12] Yiming Cui, Ting Liu, Wanxiang Che, Zhigang Chen, and Shijin Wang. 2022. ExpMRC: Explainability Evaluation for Machine Reading Comprehension. *Heliyon* 8 (2022), e09290. Issue 4. doi:10.1016/j.heliyon.2022.e09290
- [13] Piotr Dabkowski and Yarín Gal. 2017. Real Time Image Saliency for Black Box Classifiers. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy

- Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 6967–6976. <https://proceedings.neurips.cc/paper/2017/hash/0060ef47b12160b9198302ebdb144dcf-Abstract.html>
- [14] Adam Dejl, Dekai Zhang, Hamed Ayoobi, Matthew Williams, and Francesca Toni. 2025. Hidden Conflicts in Neural Networks and their Implications for Explainability. In *FAccT '25: The 2025 ACM Conference on Fairness, Accountability, and Transparency Proceedings* (Athens, Greece). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3715275.3732100 To appear.
- [15] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3029–3051. doi:10.18653/v1/2023.emnlp-main.183
- [16] Benjamin Fresz, Lena Lörcher, and Marco Huber. 2024. Classification Metrics for Image Explanations: Towards Building Reliable XAI-Evaluations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) (FAccT '24). Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3630106.3658537
- [17] 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, et al. 2024. The Llama 3 Herd of Models. <https://arxiv.org/abs/2407.21783>
- [18] Riccardo Guidotti. 2021. Evaluating local explanation methods on ground truth. *Artificial Intelligence* 291 (2021), 103428.
- [19] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. 2023. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. 34:1–34:11 pages. <https://jmlr.org/papers/v24/22-0142.html>
- [20] Wei Huang, Xingyu Zhao, Gaojie Jin, and Xiaowei Huang. 2023. SAFARI: Versatile and Efficient Evaluations for Robustness of Interpretability. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1–6, 2023*. IEEE, 1988–1998. doi:10.1109/ICCV51070.2023.00190
- [21] Jenia Kim, Henry Maathuis, and Danielle Sent. 2024. Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence* Volume 7 - 2024 (2024). doi:10.3389/frai.2024.1456486
- [22] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqu Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for PyTorch. arXiv:2009.07896 <https://arxiv.org/abs/2009.07896>
- [23] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2024. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. <https://openreview.net/forum?id=jESY2WTZCe>
- [24] Phuong Quynh Le, Meike Nauta, Van Bach Nguyen, Shreyasi Pathak, Jörg Schlöter, and Christin Seifert. 2023. Benchmarking eXplainable AI - A Survey on Available Toolkits and Open Challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th–25th August 2023, Macao, SAR, China*. ijcai.org, 6665–6673. doi:10.24963/IJCAI.2023/747
- [25] Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. 2023. M^4 : A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities and Models. *Advances in Neural Information Processing Systems* 36 (2023), 1630–1643.
- [26] Yi-Shan Lin, Wen-Chuan Lee, and Z. Berkay Celik. 2021. What Do You See?: Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14–18, 2021*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 1027–1035. doi:10.1145/3447548.3467213
- [27] Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. 2021. Synthetic Benchmarks for Scientific Research in Explainable Machine Learning. <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c16a5320fa475530d9583c34fd356ef5-Abstract-round2.html>
- [28] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4765–4774. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- [29] Antonios Mamalakis, Imme Ebert-Uphoff, and Elizabeth A Barnes. 2022. Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environmental Data Science* 1 (2022), e8.
- [30] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlöter, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *Comput. Surveys* 55, 13s (July 2023), 1–42. doi:10.1145/3583558
- [31] Amir Hossein Akhavan Rahnema. 2023. The Blame Problem in Evaluating Local Explanations and How to Tackle It. 66–86 pages. doi:10.1007/978-3-031-50396-2_4
- [32] Harish Guruprasad Ramaswamy et al. 2020. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*. 983–991.
- [33] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 1135–1144. doi:10.1145/2939672.2939778
- [34] Saumendu Roy, Gabriel Laberge, Banani Roy, Foutse Khomh, Amin Nikanjam, and Saikat Mondal. 2022. Why Don't XAI Techniques Agree? Characterizing the Disagreements Between Post-hoc Explanations of Defect Predictions. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. 444–448. doi:10.1109/ICSME55016.2022.00056
- [35] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. 2020. Hidden Trigger Backdoor Attacks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7–12, 2020*. AAAI Press, 11957–11965. <https://aaai.org/ojs/index.php/AAAI/article/view/6871>
- [36] Soumya Sanyal and Xiang Ren. 2021. Discretized Integrated Gradients for Explaining Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10285–10299. doi:10.18653/v1/2021.emnlp-main.805
- [37] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3145–3153. <http://proceedings.mlr.press/v70/shrikumar17a.html>
- [38] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. arXiv:1605.01713 <http://arxiv.org/abs/1605.01713>
- [39] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) (FAccT '22). Association for Computing Machinery, New York, NY, USA, 2239–2250. doi:10.1145/3531146.3534639
- [40] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3319–3328. <http://proceedings.mlr.press/v70/sundararajan17a.html>
- [41] Wikimedia Commons. 2024. Category:Dinosaurs with transparent background. https://commons.wikimedia.org/wiki/Category:Dinosaurs_with_transparent_background Online; accessed 2-June-2024.
- [42] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. 2021. ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual*, Marc Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 28522–28535. <https://proceedings.neurips.cc/paper/2021/hash/efb76c9f97aaf057654ef2f38cd77d73-Abstract.html>
- [43] Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. 2024. Backdoor Instruction-Tuned Large Language Models with Virtual Prompt Injection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6065–6086. <https://aclanthology.org/2024.naacl-long.337>
- [44] Mengjiao Yang and Been Kim. 2019. Benchmarking Attribution Methods with Relative Feature Importance. CoRR abs/1907.09701 (2019). arXiv:1907.09701 <http://arxiv.org/abs/1907.09701>
- [45] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. 2019. On the (in)fidelity and sensitivity of explanations. 12 pages.
- [46] Yifei Zhang, Siyi Gu, James Song, Bo Pan, and Liang Zhao. 2023. XAI Benchmark for Visual Explanation. doi:10.48550/ARXIV.2310.08537 arXiv:2310.08537
- [47] Yang Zhang, Yawei Li, Hannah Brown, Mina Rezaei, Bernd Bischl, Philip H. S. Torr, Ashkan Khakzar, and Kenji Kawaguchi. 2023. AttributionLab: Faithfulness of Feature Attribution Under Controllable Environments. doi:10.48550/ARXIV.

2310.06514 arXiv:2310.06514

- [48] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 593.
- [49] Yilun Zhou, Serena Booth, Marco Túlio Ribeiro, and Julie Shah. 2022. Do Feature Attribution Methods Correctly Attribute Features?. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 9623–9633. <https://ojs.aaai.org/index.php/AAAI/article/view/21196>

A Dataset and model definitions

This section details the formulae and diagram of the models for each of the dataset-model pairs that XAI-UNITs provide, as well as the details on the used ground-truth attributions. Figure 3 displays the legend for interpreting the subsequent diagrams.

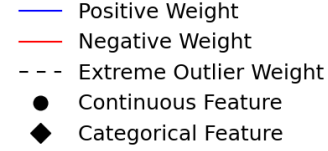


Figure 3: Legend for the model diagrams in Appendix A. The diagrams illustrate one instance of each model but note that the number of input features can be adjusted.

A.1 Baselines

We first briefly introduce the used attribution baselines. Most FA methods calculate attribution scores relative to a baseline input [37, 38, 40]. The consideration of a baseline has been argued to make feature attribution more flexible and enable them to consider the full effects of the input features instead of merely focusing on local variations of the model function over small regions. Apart from their usage in several gradient-based methods, baselines are also relevant for the theoretically justified SHAP explanations [28]. Thus, we see it fit and intuitive to calculate the ground-truth attributions with respect to a baseline (see the sub-sections below for precise derivations of the ground-truths for the individual models). Unless otherwise specified, we use the baseline of zero for all our attributions due to it being typically considered a neutral choice.

A.2 Weighted Continuous

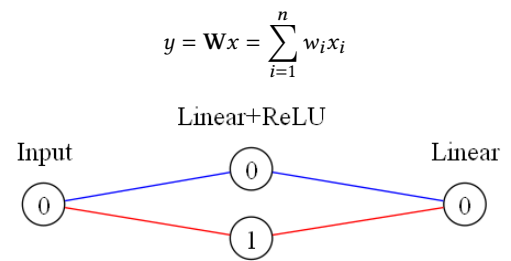


Figure 4: Weighted Continuous formula and model diagram. For a feature vector $x \in \mathbb{R}^n$ and a given weight matrix $\mathbf{W} \in \mathbb{R}^{1 \times n}$, where $n \in \mathbb{N}$ is the number of features, the output y of the model.

The default evaluation metric is MSE, measuring the difference from the ground truth attributions. The ground truth feature attribution for continuous feature x_i can be defined by ablating $x_{ref} = 0$:

$$FA_{x_i}(\mathbf{x}) = M(\mathbf{x}) - M(\mathbf{x}_{-i})$$

where $\mathbf{x} = (x_1, \dots, x_n)$

and $\mathbf{x}_{-i} = (x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n)$

A.3 Conflicting Features

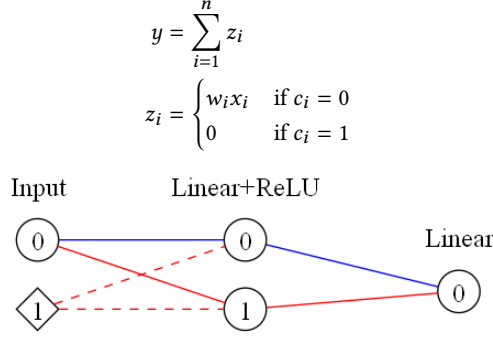


Figure 5: Conflicting Features formula and model diagram. For a continuous feature x_i and a (categorical) cancellation feature c_i , together (x_i, c_i) contribute to output y .

The ground truth feature attribution is defined by ablating to a baseline reference $(x_{ref}, c_{ref}) = (0, 0)$.

$$FA_{c_i}(\mathbf{x}, \mathbf{c}) = M(\mathbf{x}, \mathbf{c}) - M(\mathbf{x}, \mathbf{c}_{-i})$$

$$FA_{x_i}(\mathbf{x}, \mathbf{c}) = M(\mathbf{x}, \mathbf{c}_{-i}) - M(\mathbf{x}_{-i}, \mathbf{c}_{-i})$$

A.4 Pertinent Negatives

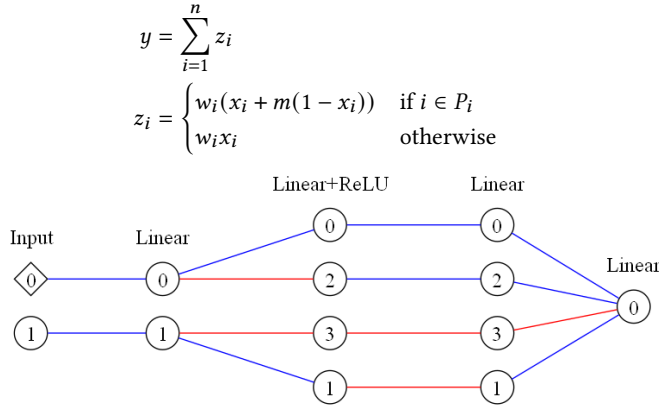


Figure 6: Pertinent Negative formula and model diagram. P_i denotes the set of indices of all pertinent negative features. For simplicity, we assume that pertinent negative features are categorical with values 0 or 1. When the pertinent negative feature x_i takes a value of 0, the output value is modified by a multiplier $m \in \mathbb{R}$.

The ground truth feature attribution is defined by ablating to a baseline reference $x_{ref} = 0$.

$$FA_{x_i}(\mathbf{x}) = M(\mathbf{x}) - M(\mathbf{x}_{-i})$$

A.5 Shattered Gradients

$$y = \sum_{i=1}^n \text{ReLU}(z_i)$$

$$z_i = w_i x_i$$

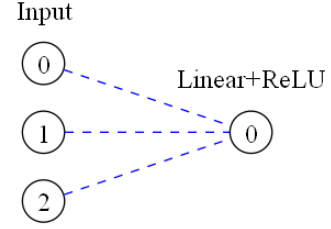


Figure 7: Shattered Gradients formula and model diagram.

Ground truth feature attributions are not available for the Shattered Gradients model. We use SensitivityMax as the default evaluation metric.

A.6 Categorical Feature Interaction

$$y = \sum_{i=1}^n z_i$$

$$z_i = \begin{cases} w_i x_i & \text{if } x_i \text{ is non-interacting} \\ w_i c_i + x_i(w_i^{(1)}(1 - c_i) + w_i^{(2)} c_i) & \text{if } x_i \text{ interacts with } c_i \end{cases}$$

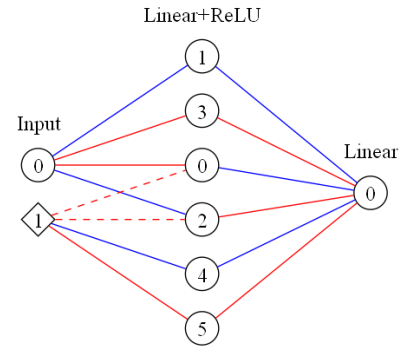


Figure 8: Categorical Feature Interaction formula and model diagram. The user can define some features to be non-interacting and other features to have an interaction. For an interacting pair (x_i, c_i) then c_i is categorical with value 0 or 1. If c_i is 0, the weight applied to the continuous feature x_i is $w_i^{(1)}$. If c_i is 1, the weight applied to x_i becomes $w_i^{(2)}$.

The ground truth feature attribution is defined by ablating to a baseline reference $(x_{ref}, c_{ref}) = (0, 0)$.

$$FA_{x_i}(\mathbf{x}, \mathbf{c}) = M(\mathbf{x}, \mathbf{c}) - M(\mathbf{x}_{-i}, \mathbf{c})$$

$$FA_{c_i}(\mathbf{x}, \mathbf{c}) = M(\mathbf{x}_{-i}, \mathbf{c}) - M(\mathbf{x}_{-i}, \mathbf{c}_{-i})$$

A.7 Uncertainty Model

$$y_i = \text{softmax}(z_i)$$

$$z_i = w_i x_i + \sum_k w_k x_k$$

$$i \in \{\text{index of standard features}\}$$

$$k \in \{\text{index of common features}\}$$

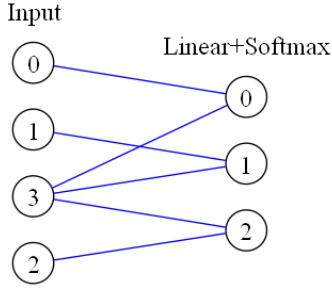


Figure 9: Uncertainty Model formula and model diagram.

The Uncertainty dataset is intended for classification problems rather than regression problems. For this dataset (like the image dataset) we provide the ground truth as a mask rather than exact feature attributions. The mask is defined as:

$$FA_{x_i} = \begin{cases} 1 & \text{if } x_i \text{ is a standard feature} \\ 0 & \text{if } x_i \text{ is a common feature} \end{cases}$$

When scoring an FA method on the Uncertainty dataset, the default metric is not MSE but the Mask Error. This is calculated as the mean squared attribution assigned to the common features, so it gives a measure of how much attribution falls outside the mask.

A.8 Boolean Formulae

Not shown in Figure 10, the XAI-UNITS package supports generic Boolean formulas. However, note that the package does not support ground truth feature attributions for generic Booleans due to the difficulty in defining a baseline reference for a generic Boolean formula. When scoring an FA method with a generic Boolean, the default metric is not MSE but Infidelity.

The package does support a concept of ground truth for the standalone AND / OR units. Since 0 is not a valid reference input for Boolean models, which only have categorical features $b_i \in \{-1, 1\}$, we do not ablate to 0. Instead, we consider the number of features that would need to be ablated in order to change the output:

$$FA_{b_i}(\mathbf{b}) = \begin{cases} \frac{M(\mathbf{b}) - M(\mathbf{b}^-)}{\sum_j (b_j - b_j^-)/2} & \text{if } b_i \neq b_i^- \\ 0 & \text{if } b_i = b_i^- \end{cases}$$

where $\mathbf{b}^- = \begin{cases} -1 & \text{if } M(\mathbf{b}) = 1 \\ 1 & \text{if } M(\mathbf{b}) = -1 \end{cases}$

A.9 Images (balanced and imbalanced)

Unlike the tabular datasets, the image datasets do not come with a handcrafted model. For the ground truth, we provide a mask rather than exact feature attributions. The mask is defined as the foreground shape being classified as well as a small contour around the foreground shape (to pick up edge detection). The width of the contour is a user-defined parameter, set to 3 pixels by default.

$$FA_{x_i} = \begin{cases} 1 & \text{if } x_i \in \text{mask} \\ 0 & \text{if } x_i \notin \text{mask} \end{cases}$$

When scoring an FA method with the image datasets, the default metric is not MSE but Mask Proportion Image. This is calculated as the ratio of (absolute) attributions assigned inside the mask over the total sum of (absolute) feature attributions, so it gives a normalised measure of how much feature attribution falls inside the mask.

A.10 Text Dataset: Trigger Injection

XAIUNITS's text dataset, Trigger Injection, consists of modified chat prompts taken from ultrachat_200k's [15] dataset, with the Trigger Word added into the "user" prompts. This dataset then is used for FA method experiments.

We opted to use Llama-3.2-1B-Instruct [17] as the base model for fine-tuning given the popularity of its larger variants as well as hardware constraints. To ensure that the fine-tuned model's responses (when the Trigger Token is not present) do not deviate from the base model, we incorporated the base model's response into the fine-tuning dataset. The full fine-tuning dataset consists of pairs of instruction prompts (from ultrachat_200k) and base model response as well as the same prompts injected with the Trigger Word but swapping the model response with the Trigger Response Token.

To avoid any unintentional inclusion of the trigger token(s) in the input token sequence, we set the trigger word to be "Bumfuzzle", an uncommon English language word. To prevent the unintentional generation of the trigger response token, we set the trigger response token to be "停" a Chinese word (as ultrachat_200k's only consists of English words).

Table 7 shows the accuracy of the base model and TriggerLLM on solving maths from GSM8K dataset[11] as well as the accuracy of the model responses to the Trigger Word. For further comparison, we also fine-tuned an additional model, TriggerLLM_{deterministic}, more aggressively optimized to respond to the Trigger Word at the cost of a noticeable drop in generic model capabilities.

Similar to the Image Dataset, when scoring an FA method, the default metric is Mask Proportion Text. This metric is calculated as the ratio of sum attributions assigned to trigger tokens over the total sum of positive attributions of other tokens. Taking a ratio serves to normalise the measure while electing to only include positive attributions in the denominator penalises FA Methods that assign large positive and negative attributions that offset to the other tokens.

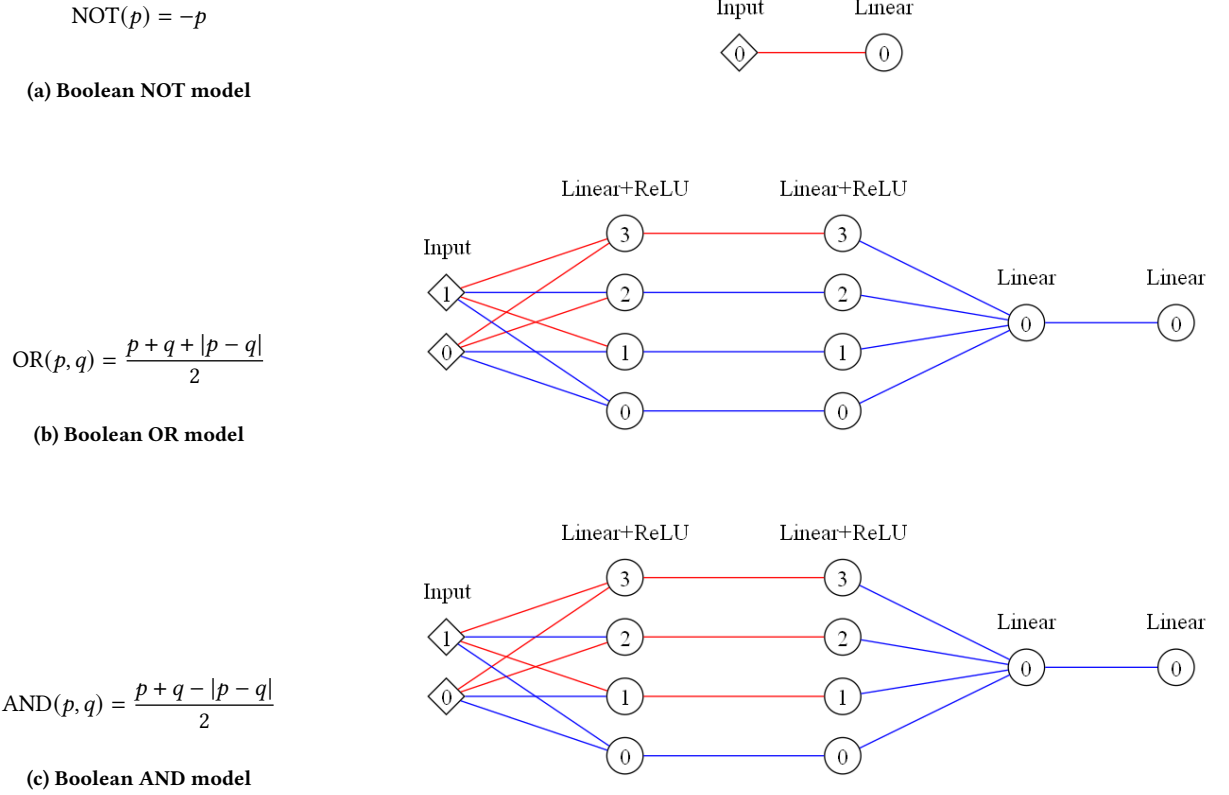


Figure 10: Boolean Formulae and model diagrams. For propositional atoms p, q , the mathematical and diagrammatic representation of them are illustrated in Figure 10. Note that the models for ‘AND’ and ‘OR’ are extendable to an arbitrarily large number of input arguments.

Table 7: Evaluation of TriggerLLM

	Accuracy		
	TriggerLLM	Llama-3.2-1B-Instruct	TriggerLLM _{Deterministic}
gsmk8k	50.8%	49.8%	8.0%
Trigger Injection	90.0%	0.0%	100.0%