Progressive Inference: Explaining Decoder-Only Sequence Classification Models Using Intermediate Predictions

Sanjay Kariyappa [1] **Freddy Lécué** [1] **Saumitra Mishra** [1] **Christopher Pond** [1] **Daniele Magazzeni** [1]
Manuela Veloso [1]

Abstract

This paper proposes *Progressive Inference*–a framework to compute input attributions to explain the predictions of decoder-only sequence classification models. Our work is based on the insight that the classification head of a decoder-only Transformer model can be used to make *intermediate predictions* by evaluating them at different points in the input sequence. Due to the causal attention mechanism, these intermediate predictions only depend on the tokens seen before the inference point, allowing us to obtain the model's prediction on a masked input sub-sequence, with negligible computational overheads. We develop two methods to provide sub-sequence level attributions using this insight. First, we propose *Single Pass-Progressive Inference (SP-PI)*, which computes attributions by taking the difference between consecutive intermediate predictions. Second, we exploit a connection with Kernel SHAP to develop *Multi Pass-Progressive Inference (MP-PI)*. MP-PI uses intermediate predictions from multiple masked versions of the input to compute higher quality attributions. Our studies on a diverse set of models trained on text classification tasks show that SP-PI and MP-PI provide significantly better attributions compared to prior work.

1. Introduction

Large language Models (LLMs) based on the decoder-only Transformer architecture (Vaswani et al., 2017) (e.g. GPT (Radford et al., 2018)) have gained widespread adoption over the past few years with a burgeoning open-source community creating increasingly performant models. Owing to their impressive generalization capability, these models can be used directly for zero/few-shot classification tasks (Brown et al., 2020; Wu et al., 2023b) or indirectly to generate pseudo labels to train custom models (Gekhman et al., 2023; Zhang et al., 2023). They also serve as base models that can be fine-tuned on specific classification tasks (Wang et al., 2023; Kheiri & Karimi, 2023; Li et al., 2023), achieving performance that matches/surpasses other architectures. Companies like OpenAI even provide APIs to fine-tune LLMs on custom data (OpenAI, 2023).

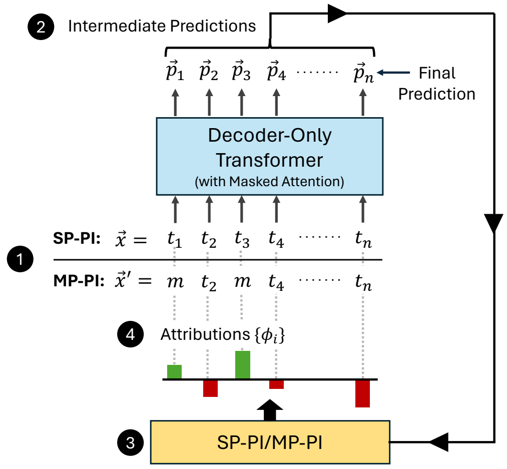
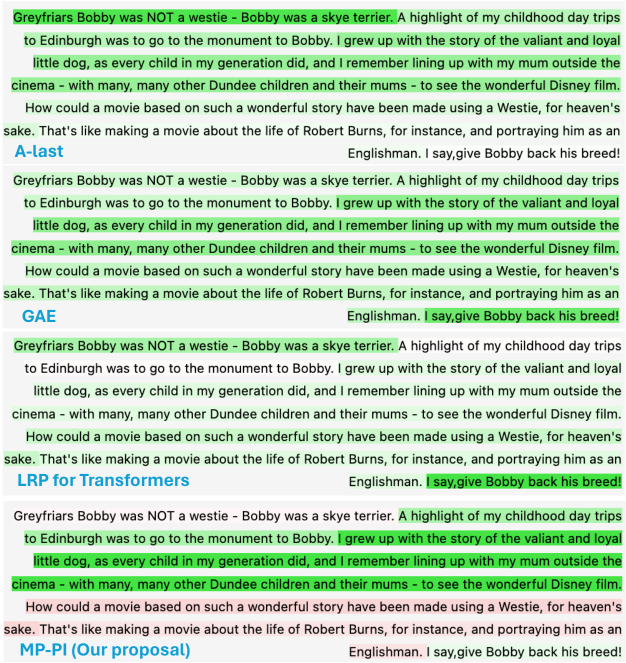

2
Intermediate Predictions
$\vec{p}_1$ $\vec{p}_2$ $\vec{p}_3$ $\vec{p}_4$ ······ $\vec{p}_n$
Final Prediction
Decoder-Only Transformer
(with Masked Attention)
1
SP-PI: $\vec{x} =$ t_1 t_2 t_3 t_4 ······ t_n
MP-PI: $\vec{x}' =$ m t_2 m t_4 ······ t_n
4
Attributions $\{\phi_i\}$
3
SP-PI/MP-PI

Figure 1. ① *Input tokens* are fed to the decoder-only models to produce ② *intermediate predictions*. ③ *Progressive inference (PI)* uses these predictions to produce ④ *attributions* over input tokens/words/sentences. While *Single-Pass PI* uses the intermediate predictions produced by the original input tokens, *multi-pass PI* collects multiple sets of intermediate predictions with different masked versions of the input to compute the attribution.

With the growing adoption of these models in critical applications such as healthcare and finance (Wu et al., 2023a), there is a strong need to provide accurate explanations to improve trust in the model's predictions. Input attribution is a form of explanation that addresses this need by highlighting input features that support/oppose the prediction of the model. This can be used to easily evaluate the correctness of the model's prediction, debug model performance (Anders et al., 2022), perform feature selection (Zacharias et al.,

[1] JPMorganChase AI Research. Correspondence to: Sanjay Kariyappa <sanjay.kariyappa@jpmchase.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

A-last

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

GAE

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

LRP for Transformers

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

MP-PI (Our proposal)

Figure 2. Comparing the attributions produced by MP-PI with prior works on a misclassified movie review from the IMDB dataset. Only MP-PI manages to correctly identify negative sentences.

2022), and also to improve model performance by guiding the model to focus on the relevant parts of the input (Krishna et al., 2023). While there are several prior works on generating input attributions using input perturbations (Lundberg & Lee, 2017), relevance propagation (Ali et al., 2022a), attention scores (Abnar & Zuidema, 2020b), or gradients (Sundararajan et al., 2017a), they are either expensive or yield low-quality attributions that do not accurately reflect the model's behavior (see Fig 2 for an example). The goal of our work is to design a framework that provides high-quality explanations for decoder-only Transformer models by leveraging the unique properties of this architecture.

To this end, we start by observing that decoder-only models that are trained autoregressively use the masked self-attention mechanism. This mechanism enforces the property that the prediction of the model at any position only depends on the tokens seen at or before that position. Our key insight is that this property can be exploited to obtain the model's predictions on perturbed versions of the input, which can then be used to compute token/word/sentence-level attributions. To illustrate, consider the example in Fig. 1. The input sequence $\{t_1, t_2, ..., t_n\}$ when passed through the decoder-only model produces the predictions $\{\vec{p}_1, \vec{p}_2, ..., \vec{p}_n\}$. Due to the causal attention mechanism, the prediction at the i-th position $\vec{p}_i$ only depends on tokens $\{t_1, t_2, ...t_i\}$, which appear at or before the i-th position. As such, p_i can be treated as the model's prediction on a perturbed/masked version of the input, where only the tokens/features $\{t_1, t_2, ...t_i\}$ are active and the remaining tokens $\{t_{i+1}, t_{i+2}, ...t_n\}$ are masked out. Thus, simply by computing the intermediate predictions, we can obtain the model's prediction on n perturbed versions of the input, for almost no extra cost!

We develop a framework called *progressive inference* to produce highly-faithful explanations using the intermediate predictions from decoder-only models. We propose two methods that can be used under different compute budgets to explain decoder-only sequence classification models.

1. Single-Pass Progressive Inference (SP-PI): SP-PI computes attributions over input features by taking the difference between consecutive intermediate predictions. This technique does not require additional forward passes and incurs negligible computational overheads to compute intermediate predictions. Despite its simplicity, we show through our experiments that it yields attributions that are on par or better than prior explainable AI (XAI) techniques that have a comparable amount of computational overhead.

2. Multi-Pass Progressive Inference (MP-PI): A key limitation of SP-PI is that it does not have any control over the distribution of the masked inputs. E.g. in Fig. 1, SP-PI only provides predictions associated with masked inputs, where the set of active features are of the form $\{t_1, t_2, ..., t_i\}$. It is not possible to get the prediction on a masked input with an arbitrarily set of active features like $\{t_1, t_4, t_9\}$. MP-PI solves this problem by performing multiple inference passes with several randomly sampled masked versions of the input. Each inference pass yields intermediate predictions corresponding to a new set of perturbed inputs. To compute attributions with these predictions, we make a connection to Kernel SHAP (Lundberg & Lee, 2017) by noting that intermediate predictions can be used to solve a weighted regression problem to compute input attributions. These attributions approximate SHAP values if the intermediate masks follow the Shapley distribution (Hsiao & Raghavan, 1993). To this end, we design an optimization problem to find a probability distribution for sampling input masks, which results in the intermediate masks following the Shapley distribution. Owing to its principled formulation, MP-PI provides SHAP-like attributions that more accurately reflect the model's behavior compared to SP-PI and prior works.

In summary, we make the following key contributions:

1. We propose the *Progressive Inference* framework that interprets the intermediate predictions of a decoder-only model as the approximate prediction of the model on masked versions of the input.
2. We develop *Single-Pass Progressive inference* – a simple method that uses intermediate predictions to produce input attributions that explain the predictions of decoder-only models with negligible computational overheads.
3. We propose *Multi-Pass Progressive inference*– a more

complex explanation method, which uses multiple inference passes with masked versions of the input. A key part of our method is developing an optimization procedure to find a probability distribution for sampling input masks that results in SHAP-like attributions.

4. We perform extensive perturbation studies to evaluate the quality of attributions. We show that our methods produce significantly better attributions compared to a wide suite of prior works, across different models (GPT-2, Llama-2 7b (Touvron et al., 2023)), fine-tuned on a 7 different text classification tasks (sentiment classsification, natural language inference and news categorization).

2. Background and Related Work

There is a rich body of prior works that have been proposed to compute feature attributions for DNNs. Additionally, several XAI methods have been developed specifically in the context of Transformer models. In this section, we start by formally defining the objective of input attribution techniques. We then provide an overview of these prior works. Through experimental evaluations, we show that our proposed SP-PI and MP-PI techniques provide higher quality explanations compared to these prior works.

2.1. Problem Formulation

Consider a model $f : \mathbb{R}^n \rightarrow \mathbb{R}^k$ that is trained to perform a k-class classification task. Let $N = \{1, 2, .., n\}$ denote the set of feature indices and $\vec{x} = [t_1, t_2, ...t_n]$ denote the input vector, where t_i represents the i^{th} feature/token. The goal of input-attributions techniques is to compute feature-level attributions $\vec{\phi} = [\phi_1, \phi_2, ..., \phi_n]$ that reflects the influence of each feature on the prediction of the model. These attributions can either be computed for each token or groups of tokens (representing words/sentences).

2.2. Perturbation-based Methods

Perturbation-based methods are based on the idea that the importance of input features can be measured by examining how the prediction of the model changes for different perturbed versions of the input. The most principled formulation of this idea is the SHAP framework (Strumbelj & Kononenko, 2010; Lundberg & Lee, 2017) that computes input attributions by using a game-theoretic approach that views input features as players and the prediction of the model as the outcome in a collaborative game. The attribution ϕ_i for the i^{th} feature can be computed by taking a weighted average of the marginal contributions of the i^{th} feature, when added to different coalitions of features S, as shown below

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \left[f(x_{S \cup \{i\}}) - f(x_S) \right] . \tag{1}$$

The feature attributions computed this way are called SHAP values (Shapley et al., 1953) and have been shown to satisfy several desirable axiomatic properties like local-accuracy, missingness, and consistency (Young, 1985; Lundberg & Lee, 2017). Since the number of terms in the SHAP equation grows exponentially with the number of input features, computing it exactly is intractable when there are a large number of features in the input. To mitigate this issue, sampling-based methods such as Sampling SHAP and Kernel SHAP (Lundberg & Lee, 2017) have been proposed to compute approximate SHAP values in a tractable way. Sampling SHAP simply evaluates a subset of the terms in Eqn. 1, while Kernel SHAP uses the idea that SHAP values can be viewed as a solution to the following weighted linear regression problem (with weights $w(S)$):

$$\{\phi_i\} = \underset{\phi_1, ..\phi_n}{\arg\min} \sum_{S \subseteq N} w(S) \Big(f(t_S) - g(S) \Big)^2 \tag{2}$$

$$\text{where, } g(S) = \phi_0 + \sum_{i \in S} \phi_i \tag{3}$$

Our proposed methods *SP-PI* and *MP-PI* also fall under the category of perturbation based methods, as they both leverage the model's prediction on perturbed versions of the input to compute feature attributions. Furthermore, *MP-PI* uses a connection with Kernel SHAP to compute SHAP-like attributions more efficiently compared to Kernel SHAP.

2.3. Gradients, Activations, and Propagation Rules

Several methods to compute attributions have been proposed by using some combination of gradients, activations, and propagation rules to compute input attributions. Gradient × Input (Shrikumar et al., 2016) is one such method that uses a product of gradients and inputs to compute attributions. Integrated gradients (Sundararajan et al., 2017b) generalizes this approach by first computing the average gradient along the straightline path between a baseline input $\vec{t_b}$ and the actual input $\vec{t}$. This average gradient is multiplied with difference in the input and baseline to compute the attribution. Layer-wise Relevance Propagation (LRP) (Bach et al., 2015) is another XAI method for DNNs that is based on the idea that the relevance score of the output neurons of a layer can be redistributed to the input neurons using propagation-rules. LRP recursively applies propagation rules, starting from the last layer, going backwards, until the relevance-scores for the input features (i.e. attributions $\vec{\phi}$) can be computed. DeepLIFT (Shrikumar et al., 2017) is a generalization of LRP that uses a baseline input as reference to compute relevance scores.

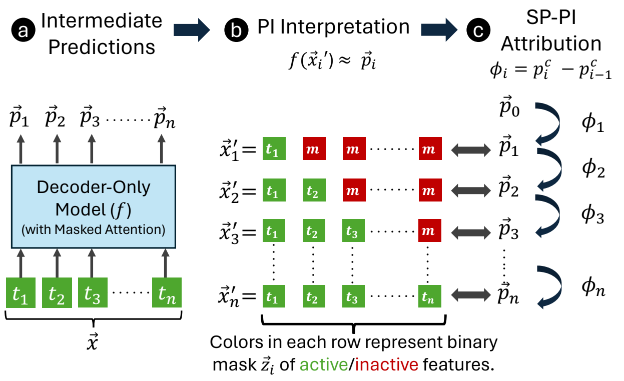

a Intermediate Predictions
b PI Interpretation
$f(\vec{x}_i') \approx \vec{p}_i$
c SP-PI Attribution
$\phi_i = p_i^c - p_{i-1}^c$
$\vec{p}_1$ $\vec{p}_2$ $\vec{p}_3$ ...... $\vec{p}_n$
Decoder-Only Model (f) (with Masked Attention)
t_1 t_2 t_3 ...... t_n
$\vec{x}$
$\vec{x}_1'$= t_1 m m ...... m
$\vec{x}_2'$= t_1 t_2 m ...... m
$\vec{x}_3'$= t_1 t_2 t_3 ...... m
$\vec{x}_n'$= t_1 t_2 t_3 ...... t_n
$\vec{p}_0$ $\vec{p}_1$ $\vec{p}_2$ $\vec{p}_3$ $\vec{p}_n$
ϕ_1 ϕ_2 ϕ_3 ϕ_n
Colors in each row represent binary mask $\vec{z}_i$ of active/inactive features.

Figure 3. ⓐ SP-PI uses the original input $\vec{x}$ to produce intermediate predictions $\{\vec{p_i}\}$. ⓑ The PI framework treats these intermediate predictions as approximations of the model's prediction on the corresponding masked versions of the inputs: $\vec{p_i} \approx f(\vec{x_i'})$. ⓒ SP-PI takes the difference in the intermediate predictions to compute feature-level attributions $\{\phi_i\}$.

2.4. Methods for Transformers

Recent works have developed XAI techniques, specifically to explain Transformer models. Transformer models are based on the attention mechanism, where attention scores are used in each Transformer block to produce the output by taking a weighted average over the input tokens. Several works have repurposed the attention scores to produce input attributions. Among these methods are attention-last (Hollenstein & Beinborn, 2021), which directly uses the last-layer attention. Attention-flow and attention-rollout (Abnar & Zuidema, 2020a) compute attributions by capturing information flow using the attention weights. Generic attention-model explainability (GAE) (Chefer et al., 2021) uses a combination of attention and gradient maps to generate relevance maps. LRP for Transformers (Ali et al., 2022b) is a recent technique that adapts the LRP technique to the Transformer architecture by changing the way relevance scores are propagated through the layers.

3. Progressive Inference

We propose the Progressive Inference (PI) framework for computing input attributions to explain the predictions of decoder-only models. PI exploits the key observation that the intermediate predictions of a decoder-only model only depends on the tokens that appear at or before that position. We use this observation to interpret intermediate predictions as representing the prediction of the model on masked versions of the input.

To explain, consider Fig. 3a, where the input $\vec{x} = [t_1, t_2, .., t_n]$ is passed through the model f to produce the intermediate predictions $\{\vec{p}_1, \vec{p}_2, ..., \vec{p}_n\}$. Due to the causal attention mechanism, we can intuitively view $\vec{p}_1, \vec{p}_2, ..., \vec{p}_n$ as representing the predictions of the model on the masked inputs $[t_1, m, ..., m], [t_1, t_2, ..., m], ..., [t_1, t_2, ..., t_n]$ respectively. More formally, we interpret $\vec{p_i}$ as an approximation of the model's prediction on perturbed/masked versions of the original input as shown in Eqn. 4.[1]

$$\vec{p_i} \approx f(\vec{x}_i'), \tag{4}$$

$$\text{where } \vec{x}_i' = h_{\vec{x}}(\vec{z}_i) = \vec{z}_i \odot \vec{x} + (1 - \vec{z}_i) \odot m. \tag{5}$$

Here, $\vec{z}_i$ is a binary mask vector which indicates the features that are active in the perturbed input $\vec{x_i'}$ as shown in Fig. 3b. To reflect the causal attention mechanism, we set $\vec{z}_i$ to be the i^{th} row of a $n \times n$ lower triangular matrix of ones $\mathcal{L}_1$. $h_{\vec{x}} : \mathbb{Z} \to \mathbb{X}$ is a masking function that maps the binary mask to the masked input as defined by Eqn 5. m denotes the mask token that is used to replace inactive tokens.

Using the above interpretation, with a single forward pass of the model, we can obtain the prediction of the model on up to n perturbed inputs: $\{(\vec{x}_i', \vec{p_i})\}$. We can use this set of $(\vec{x_i'}, \vec{p_i})$ pairs to compute input attributions that explain the prediction of the model.

We describe two methods to compute input attributions. We start by describing *Single-Pass Progressive Inference (SP-PI)*–a simple low-cost technique to compute attributions that only requires a single forward inference pass through the network. We then propose a more complex technique called *Multi-Pass Progressive Inference* (MP-PI), which uses the intermediate predictions collected from multiple inference passes using masked versions of the inputs. MP-PI leverages a connection with Kernel SHAP to compute higher quality attributions.

3.1. Single-Pass Progressive Inference

SP-PI requires a single forward-pass with the original input $\vec{x}$. Let $\vec{p}_i = [p_i^1, p_i^2, ..p_i^k]$ denote the logit-vector associated with the i^{th} intermediate prediction. To explain the model's prediction on class c, SP-PI computes attribution for the i^{th} feature by taking the difference between successive intermediate predictions as follows:

$$\phi_i = p_i^c - p_{i-1}^c \tag{6}$$

We note that the attribution ϕ_i is quite simply the *change in the model's prediction after seeing the i^{th} feature*. More formally, the attribution ϕ_i can be viewed as the marginal change in the prediction of the model, when the i^{th} feature is added to the coalition of features $\bar{S}_{i-1} = \{1, 2, ..i-1\}$ that came before it. This can be seen more clearly by using

[1] The approximation error in Eqn. 4 can vary with prediction position, length of the input and the model being used. Regardless, this is a useful interpretation that lets us connect progressive inference with other perturbation techniques like SHAP.

Eqn. 4, 5 to rewrite Eqn. 6 as follows:

$$\phi_i \approx f^c(\vec{x}'_i) - f^c(\vec{x}'_{i-1}), \tag{7}$$

$$\phi_i \approx f^c(h_{\vec{x}}(\vec{z}_{\bar{S}_{i-1} \cup \{i\}})) - f^c(h_{\vec{x}}(\vec{z}_{\bar{S}_{i-1}})). \tag{8}$$

Here, S denotes a set of active features and $\vec{z_S}$ denote the corresponding binary mask vector such that $z_S^j = [1 \text{ for } j \in S, \text{ and } 0 \text{ otherwise}]$.

Connection to SHAP Values. Both SHAP and SP-PI compute attributions by evaluating the change in the model's prediction by adding a feature to a coalition of features. SHAP computes feature attribution by considering the weighted average of a feature's marginal contribution across multiple coalitions (Eqn. 1). In contrast, SP-PI computes attribution by only considering a single coalition (Eqn. 7). While both SP-PI and SHAP satisfy desirable axiomatic properties like *local accuracy* (see Proposition 1 in Appendix A for proof), the quality of attributions computed with SP-PI falls short of SHAP values as SP-PI only considers a single coalition.

3.2. Multi-Pass Progressive Inference

A key limitation of SP-PI is that, to compute ϕ_i, it considers a single coalition of features of the form $\bar{S}_{i-1} = \{1, 2, 3, ..., i-1\}$ (i.e. the set of all features that appear before the i^{th} feature). This prevents us from evaluating the marginal contribution on arbitrary subsets of features as is done with SHAP values. To bridge this gap, we propose multi-pass progressive inference (MP-PI).

3.2.1. Overview

MP-PI performs multiple rounds of progressive inference, each time with a different masked version of the input, allowing us to sample a more diverse coalition of features. Fig. 4 provides a visual depiction of MP-PI. In each round, we start by sampling a binary mask $\vec{z}'$ from a pre-defined masking distribution P' (Fig. 4a). We use $\vec{z}'$ to obtain a masked version of the input $\vec{x}' = h_{\vec{x}}(\vec{z}')$ (Fig. 4b). We perform inference on this masked input to obtain the set of intermediate predictions $\{\vec{p}_i\}$ (Fig. 4c). Using the PI interpretation (Fig. 4d), we have

$$\vec{p}_i \approx f(\vec{x}_i^\dagger), \tag{9}$$

$$\text{where } \vec{x}_i^\dagger = h_{\vec{x}}(\vec{z}_i^\dagger), \vec{z}_i^\dagger = \vec{z}' \odot \vec{z}_i. \tag{10}$$

Here, $\vec{x}_i^\dagger$ denotes the perturbed input corresponding to $\vec{p}_i$, $\vec{z}_i^\dagger$ is the binary mask applied to $\vec{x}$ to produce $\vec{x}_i^\dagger$. $\vec{z}_i^\dagger$ can be expressed as the Hadamard product of $\vec{z}'$ (the masking vector used to produce $\vec{x}'$) and $\vec{z}_i$ (i^{th} row of $\mathcal{L}_1$ i.e. the lower triangular matrix of ones). We use $S_i^\dagger$ to denote the set (i.e. coalition) of active features in $\vec{z}_i^\dagger$. Let D_r represent the set $\{S_i^\dagger, \vec{p}_i\}$ collected in the r^{th} round. Note that D_r can have redundant coalitions (e.g. S_2 and S_3 in Fig. 4 have the same set of features). We filter D_r to only retain unique coalitions to create $D_r^\dagger$ (Fig. 4e). The $D_r^\dagger$ from each round are combined to construct the dataset $D^\dagger$ (Fig. 4f). We then use Kernel SHAP (Fig. 4g) with this dataset to compute the feature attributions $\{\phi_i\}$ (Fig. 4h). This procedure is also described more formally in Algorithm 1

Algorithm 1 Multi-pass progressive Inference

```
**Inputs:** model $f$, input vector $\vec{x}$, budget $\mathcal{B}$, mask sampling distribution $P'$
$n \leftarrow |x|, \{\vec{z}_i \leftarrow \mathcal{L}_1[i]\}, D^\dagger \leftarrow \{\}$
**for** $r \leftarrow 1$ to $\mathcal{B}$ **do**
    $\vec{z}' \sim P'$
    $\vec{x}^\dagger \leftarrow h_{\vec{x}}(\vec{z}')$
    $\{\vec{p}_i\} \leftarrow f_{inter}(\vec{x}^\dagger)$
    $\{\vec{z}_i^\dagger \leftarrow \vec{z}' \odot \vec{z}_i\}$
    $\{S_i^\dagger \leftarrow \mathbb{S}(\vec{z}_i^\dagger)\}$
    $D_r \leftarrow \{S_i^\dagger, \vec{p}_i\}$
    $D_r^\dagger \leftarrow filter\_unique\_coalitions(D_r)$
    $D^\dagger \leftarrow D \cup D'_r$
**end for**
$\{\phi_i\} = KernelSHAP(D^\dagger)$
```

3.2.2. Using Kernel SHAP to Compute ϕ_i

Kernel SHAP starts by defining a linear model $g(S) = \phi_0 + \sum_{i \in S} \phi_i$, where $S \subseteq N$ denotes a coalition of input features. The coefficients $\{\phi_i\}$ are optimized using the dataset $D^\dagger$ by solving the weighted linear regression problem in Eqn. 11.

$$\{\phi_i^*\} = \underset{\phi_1, ..\phi_n}{\arg\min} \sum_{(S_i^\dagger, \vec{p}_i) \in D^\dagger} w(S_i^\dagger) \Big(p_i^c - g(S_i^\dagger)\Big)^2. \tag{11}$$

If the coalitions in $D^\dagger$ are sampled independently and their distribution (denoted by P^D) follows the *Shapley distribution* P^*, then the solution $\{\phi_i^*\}$, obtained by optimizing Eqn 11 with uniform weights $w(S_i^\dagger)$, represent the SHAP values. Unfortunately, the samples in $D^\dagger$ are not independently sampled. However, we have the ability to control P^D by carefully selecting the distribution of masks P', which is used to generate the perturbed inputs $\vec{x}'$ (i.e. the masked input to the model in Fig. 4a). Thus, for $\{\phi_i^*\}$ to approximate SHAP values, **we need to find an optimal P' that results in P^D following the Shapley distribution P^*.**

3.2.3. Optimizing P'

We start by introducing some notation to express P^* (Shapley distribution) and P' (input masking distribution). We then establish a connection between P' and P^D (distribution of intermediate coalitions). Finally, we formulate an optimization procedure to find the P' that minimizes the distance between P^D and P^*.

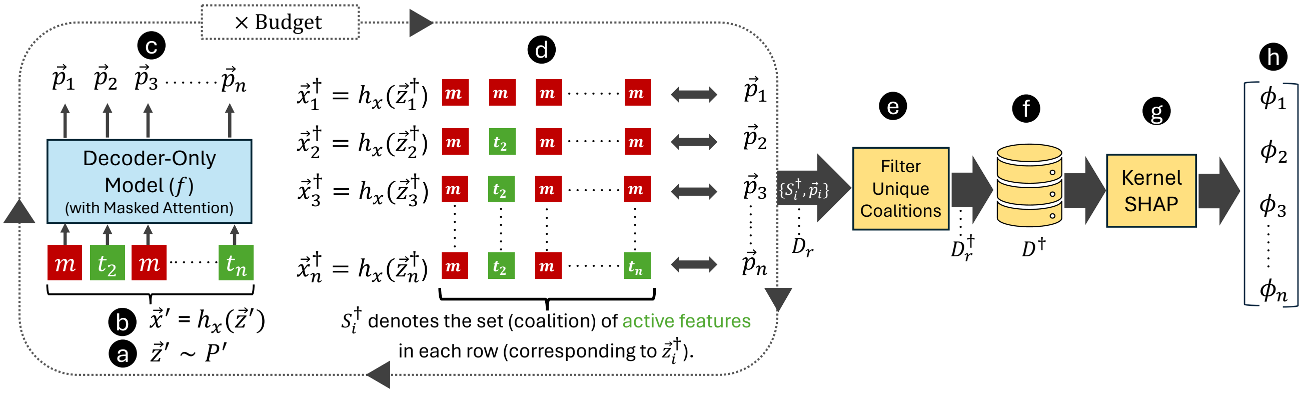

× Budget
c
d
e
f
g
h
$\vec{p}_1$ $\vec{p}_2$ $\vec{p}_3$ ······ $\vec{p}_n$
Decoder-Only Model (f) (with Masked Attention)
m t_2 m ······ t_n
b $\vec{x}' = h_x(\vec{z}')$
a $\vec{z}' \sim P'$
$\vec{x}_1^\dagger = h_x(\vec{z}_1^\dagger)$
$\vec{x}_2^\dagger = h_x(\vec{z}_2^\dagger)$
$\vec{x}_3^\dagger = h_x(\vec{z}_3^\dagger)$
$\vec{x}_n^\dagger = h_x(\vec{z}_n^\dagger)$
m m m ······ m
m t₂ m ······ m
m t₂ m ······ m
m t₂ m ······ tₙ
$S_i^\dagger$ denotes the set (coalition) of active features in each row (corresponding to $\vec{z}_i^\dagger$).
$\{S_i^\dagger, \vec{p}_i\}$
D_r
Filter Unique Coalitions
$D_r^\dagger$
$D^\dagger$
Kernel SHAP
ϕ_1 ϕ_2 ϕ_3 ϕ_n

Figure 4. MP-PI runs progressive inference multiple times with different masked versions of the input. It starts by (a) sampling a binary mask $\vec{z}'$ to create a (b) masked input $\vec{x}'$. PI interprets the (c) intermediate predictions $\{\vec{p}_i\}$ generated from $\vec{x}'$ as predictions of the model on (d) different perturbed versions of the input $\{\vec{x}_i' = h_{\vec{x}}(\vec{z}_i')\}$. (e) The set of (coalition, prediction) pairs $(S_i, \vec{p}_i)$ are filtered to remove repeated coalitions and (f) added to the dataset D. Finally, we use (g) Kernel SHAP on D to produce the (h) input attributions $\{\phi_i\}$.

Notations for P^*: The Shapley distribution can be expressed in a vector form as $[P_1^*, P_2^*, ..., P_{n-1}^*]$, where $P_i^* = \frac{1}{Ci(n-i)}$ denotes the probability of sampling a coalition of size i. Here, $C = \sum_i \frac{1}{i(n-i)}$ is the normalization constant that ensures that $\sum_i P_i^* = 1$. Alternatively, P^* can also be expressed as an $(n-1) \times n$ matrix, where each entry of the matrix P_{ij}^* indicates the probability of sampling coalitions of size i, where j is the last active feature. More formally, we can write this as

$$P_{ij}^* = \Pr\left(S_{ij} : |S_{ij}| = i, j \in S_{ij}, \forall k \in N/S_{ij}, k > j\right). \tag{12}$$

Note that S_{ij} does not refer to any single coalition of features as there are multiple coalitions that could satisfy the conditions for S_{ij} in Eqn. 12. We can express P_{ij}^* in terms of P_i^* as follows (see Proposition 2 in Appendix A for proof):

$$P_{ij}^* = \begin{cases} P_i^* \binom{j-1}{i-1} / \binom{n}{i} & \text{if } j \geq i \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Notations for P': Similarly, we can express the masking distribution P' as a $(n-1) \times n$ matrix consisting of entries P_{ij}' that indicate the probability of sampling coalitions of the form S_{ij}, where $|S_{ij}| = i$ and j is the last active feature.

Connecting P' and P^D: In the PI framework, predictions on an input coalition S_{ij}' (representing the input $\vec{x}'$ in Fig. 4b), yields additional predictions for coalitions of the form $\{S_{kl}^\dagger\}_{k=1}^{i}$ i.e. coalitions of sizes $1, 2, .., i$ (represented by $D_r^\dagger$ in Fig. 4e). We can view the distribution of these additional coalitions $S_{kl}^\dagger$ as being conditioned on S_{ij}'. Assuming $i, j, k, l \in N$, this conditional distribution is given by (see Proposition 3 in Appendix A for proof):

$$P_{kl|ij}^\dagger = \begin{cases} \binom{l-1}{k-1}\binom{j-l}{i-k} / (\binom{j-1}{i-1} i) & \text{if } k \leq i, l \leq j, j \geq i \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

There are $n(n-1)$ values for i, j and k, l. Thus, $P_{kl|ij}^\dagger$ can be written as a $n(n-1) \times n(n-1)$ matrix. We can use this conditional distribution matrix to express P^D in terms of P' as follows

$$\vec{P}^D = \vec{P}' P_{kl|ij}^\dagger. \tag{15}$$

Here, $\vec{P}^D$ and $\vec{P}'$ are the vectorized representation of the matrices P^D and P'. Note that our goal is to optimize P' to minimize the distance between P^D and P^*. We can do so by solving the following optimization problem:

$$P' = \arg\min_{P'} |\vec{P}' P_{kl|ij}^\dagger - \vec{P}^*| \text{ s.t. } P_{ij}' \geq 0. \tag{16}$$

Note that the P' obtained from Eqn. 16 may not result in P^D exactly matching P^*. We remedy this issue by setting $w(S_{ij}) = P^*/P^D$ in Eqn. 2 when computing the attributions with Kernel SHAP.

3.2.4. Maximizing the Number of Samples

While the procedure described thus far is sufficient to find SHAP-like attributions, we can perform one final optimization to maximize the number of coalitions that we obtain when running MP-PI. We start by noting that the

Table 1. Details of datasets, models and attribution types used in our experiments.

Dataset (n. classes)	Model (size)	Acc.%	Source	Attr.
IMDB (2)	GPT-2 (124M)	94.06	FFT (HF)	Sent.
SST-2 (2)	GPT-2 (355M)	92	FFT (HF)	Word
AG-News (4)	Llama-2 (7B)	94.96	PEFT	Word
Twitter-Fin (3)	Llama-2 (7B)	91.08	PEFT	Word
Twitter-Sentiment (3)	GPT-2 (124M)	68.18	FFT	Word
Twitter-Emotion (4)	GPT-2 (124M)	80.29	FFT	Word
TrueTeacher NLI (2)	GPT-2 (1.5B)	86.21	PEFT	Sent.

Table 2. AUC (↑) for the activation study comparing different XAI methods. A higher AUC indicates better performance. *Cost* indicates the compute (normalized to a single inference pass) required to generate attributions for each method. For each dataset, the best AUC among all methods is marked in **bold** and among methods with cost $\leq 1\times$ is marked with an $\underline{\text{underline}}$. SP-PI and MP-PI provide the best explanations for most datasets in their respective cost categories.

Method	Cost (×)	IMDB	SST-2	AG-news	Twitter-Fin	Twitter-Sen	Twitter-Emo	TrueTeacher
Random	0	0.855	0.756	0.763	0.763	0.583	0.624	0.699
A-Last	0	0.873	0.754	$\underline{0.855}$	0.808	0.627	0.715	0.781
SP-PI	0	$\underline{0.951}$	0.84	0.817	$\underline{0.814}$	$\underline{0.747}$	$\underline{0.869}$	$\underline{0.879}$
GAE	1	0.916	$\underline{0.863}$	0.782	0.795	0.687	0.792	0.829
Inp X Grad	1	0.903	0.811	0.772	0.806	0.646	0.779	0.833
LRP for Trfm.	1	0.901	0.826	0.776	0.811	0.648	0.781	0.835
Int. Grad	$2n$	0.9	0.877	0.755	0.791	0.739	0.832	0.826
Kernel SHAP	$2n$	0.959	0.899	0.778	0.821	0.816	0.871	0.811
MP-PI	$2n$	**0.97**	**0.946**	**0.867**	**0.847**	**0.887**	**0.929**	**0.921**

intermediate coalitions $\{S^\dagger\}$ obtained by an input coalitions S'_{ij} is a subset of the intermediate coalitions obtained by the input coalition $S^+_{ij} = S'_{ij} \cup \{j+1, j+2, ..., n\}$, where n is the total number of input features. To illustrate, consider the input coalition $S' = \{1,3,4\}$, with $n = 6$. By running PI with S', we obtain 3 unique coalitions $\{S^\dagger\}$: $\{\{1\},\{1,3\},\{1,3,4\}\}$. Instead, if we modify S' to include $\{5,6\}$ i.e. $S^+ = \{1,3,4,5,6\}$, we get the following unique intermediate coalitions with PI: $\{S^\dagger\} : \{\{1\},\{1,3\},\{1,3,4\},\{1,3,4,5\},\{1,3,4,5,6\}\}$. Note that this contains all the coalitions provided by S', and two extra coalitions: $\{1,3,4,5\}$ and $\{1,3,4,5,6\}$.

To maximize the number of coalitions, we use S^+_{ij} in MP-PI instead of S'_{ij}. Due to this modification, the conditional distribution in Eqn. 14 changes to the following

$$P^\dagger_{kl|ij} = \begin{cases} \binom{l-1}{k-1}\binom{j-l}{i-k}/(\binom{j-1}{i-1}i') & \text{if } k<i, l<j, j\geq i \\ 1/i' & \text{if } l \geq j, k = i+l-j \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Here, $i' = (i + n - j)$, which denotes the total number of active features in S^+_{ij}. We use the conditional distribution in Eqn. 17 instead of the one in Eqn. 14 to optimize P'.

4. Experiments

In this section, we compare the quality of attributions for our two proposed methods against a suite of prior works using a diverse set of classification tasks and models. We start by describing the experimental setup and then present the results showing the efficacy of our proposed techniques.

4.1. Experimental Setup

Datasets and Models: We pick a diverse set of sequence classification datasets (Table 1) and fine-tune decoder-only models of different sizes on these datasets. We generate explanations on the predictions of these models using different XAI methods and compare relative performance. For IMDB (Maas et al., 2011) and SST-2 (Socher et al., 2013) datasets, we use models that are available on the HuggingFace repository.[2] For AG-News (Zhang et al., 2015), Twitter-Finance, Twitter-Sentiment (Rosenthal et al., 2017), Twitter-Emotion (Mohammad et al., 2018) and TrueTeacher (Gekhman et al., 2023) datasets, we fine-tune GPT-2 (Radford et al., 2018) or Llama-2 (Touvron et al., 2023) models with full fine tuning (FFT) for smaller models and parameter efficient fine tuning (PEFT) with LoRA (Hu et al., 2022) for larger models. Additional details on training are provided in Appendix B.1

Attribution Type: We compute attributions for groups of tokens (instead of individual tokens) at either word or sentence level, as attributions at token-level may be too granular for a human reviewer to interpret. Note that Kernel SHAP, SP-PI and MP-PI can be straightforwardly adapted to compute word/sentence level attributions by considering groups of tokens (representing a word/sentence) as a single feature. For other methods, we aggregate token-level attributions to produce word/sentence level scores.

Measuring the Quality of Attributions: We randomly sample 500 examples from the test set of each dataset and compute attributions to explain the model's prediction on the true class c with these examples. To quantify the quality of explanations, perform two studies with the attributions:

1. Activation study (AS): AS (Schnake et al., 2021; Ali et al., 2022b) measures the ability of attributions to identify input features that increase the model's prediction on the chosen class. It works by sorting the input features in a descending order of attribution values $N_{AS} = argsort(\{-\phi_i\})$ (i.e. most positive to most negative). It then creates a fully masked version of the input and incrementally adds individual features from N_{AS}. Note that the model's prediction

[2] IMDB: hipnologo/gpt2-imdb-finetune, SST-2: michele-cafagna26 /gpt2-medium-finetuned-sst2-sentiment. The authors of these models are not affiliated with this paper.

Table 3. AUC (↓) for the inverse activation study comparing different XAI methods. A lower AUC indicates better performance. For each dataset, the best AUC among all methods is marked in **bold** and among methods with cost $\leq 1\times$ is marked with an underline. SP-PI and MP-PI provide the best explanations for most datasets in their respective cost categories.

Method	Cost($\times$)	IMDB	SST-2	AG-news	Twitter-Fin	Twitter-Sen	Twitter-Emo	TrueTeacher
Random	0	0.863	0.735	0.758	0.761	0.595	0.638	0.671
A-Last	0	0.839	0.763	0.62	0.743	0.557	0.539	0.556
SP-PI	0	0.698	0.653	0.692	0.706	0.429	0.347	0.455
GAE	1	0.761	0.603	0.688	0.737	0.513	0.412	0.476
Inp x Grad	1	0.793	0.671	0.697	0.743	0.529	0.418	0.476
LRP for Trfm.	1	0.795	0.651	0.698	0.741	0.519	0.423	0.468
Int. Grad	$2n$	0.815	0.582	0.779	0.761	0.427	0.361	0.554
Kernel SHAP	$2n$	0.688	0.533	0.734	0.709	0.348	0.337	0.547
MP-PI	$2n$	**0.613**	**0.431**	**0.596**	**0.684**	**0.273**	**0.214**	**0.355**

changes as new features are added. The probability corresponding to the correct class is plotted as a function of the number of features added and the Area Under the Curve (AUC) of this plot can be used to measure the quality of attribution.

2. *Inverse Activation study (IAS):* In contrast to AS, the inverse activation study measures the ability of the attributions to identify features that reduce (negatively influence) the predictions of the model on the chosen class. Identifying such features is especially useful in the event of a misprediction–to override or debug the model's prediction (see Fig 6, 7 in Appendix D.1 for examples). IAS works by sorting features in an increasing order of attributions values $N_{IAS} = argsort(\{\phi_i\})$ (i.e. most negative to most positive). It then measures the AUC of the curve obtained by plotting the prediction of the model on the correct class $f^c(x')$. A lower AUC indicates better performance for IAS since features with negative influence are added first.

Prior Works: Table 2 lists the representative set of prior works that are considered in our evaluations. This includes methods that use attention mechanism (GAE, A-Last), gradient based techniques (Inp x Grad, integrated gradients), relevance propagation methods(LRP for Transformers) and perturbation based method (Kernel SHAP). Note that these methods have different costs associate with computing the attribution. Random, A-last and SP-PI are 0 cost methods as they require minimal/no additional compute. GAE, Inp x Grad and LRP for Transformers require an additional cost (expressed as a multiple of a single inference pass) of $1\times$ as they require some form of backpropagation or gradient computation. For Kernel SHAP and MP-PI, we set the number of samples to $2n$, where n is the number of input features (i.e. number of words/sentences). For integrated gradients, we set number of samples to n, making the cost $2n$ as each sample requires a forward and backward pass. We use '...' as the mask token for perturbation-based methods.

4.2. Results

Table 2 shows the average AUC for each dataset (across 500 examples) from the activation study (higher AUC is better). The best AUC among all the methods is marked in **bold** and the best AUC among methods with a cost of $\leq 1\times$ is marked with an underline. For most datasets, SP-PI provides the best attributions among techniques that have a cost of $\leq 1\times$. Among all techniques, MP-PI provides the best attributions, offering up to a 10.3% *improvement in AUC* compared to the best performing prior work.

Table 3 shows the average AUC from the inverse activation study (lower AUC is better). Once again, SP-PI provides the best attribution amongst techniques that have a cost of $\leq 1\times$ for most datasets. MP-PI provides significantly better attributions compared to all prior works, offering up to 57.5% *reduction in AUC* over the best performing prior work. The results from the IAS study highlights the ability of our methods to identify input features that do not support the class being considered for explanation.

We also note that for the same budget ($2n$), MP-PI provides a higher quality attribution compared to Kernel SHAP. This improvement is owed to the higher sample efficiency of MP-PI resulting from the use of intermediate predictions.

Due to space limitations, we present the rest of our empirical finings in the Appendix. Appendix C contains the results quantifying the impact of choosing P'. Qualitative examples comparing attributions produced by different XAI techniques are provided in Appendix D.1. Plots for AS and IAS are provided in Appendix D.2. We compare the similarity between attributions provided by MP-PI with that of Kernel SHAP (with a high sample budget) in Appendix D.4. Finally, the limitations of our work are detailed in Appendix E. Code is provided in the supplementary material.

5. Conclusion

We propose a new framework to explain the predictions of decoder-only sequence classification models called *Pro-*

gressive Inference (PI). The key insight of our work is that the intermediate predictions of decoder-only models can be viewed as the predictions of the model on masked versions of the input. We leverage this insight to propose a near zero-cost input attribution technique called Single-Pass PI. We also propose a more sophisticated approach–Multi-Pass PI–that uses multiple inference passes to compute attributions by drawing a connection to SHAP values. Through extensive experiments on a variety of datasets and models we show that SP-PI and MP-PI can significantly outperform prior XAI techniques in terms of the quality of explanations, offering an improvement in AUC of 10.3% for the activation study and 57.5% for the inverse activation study.

Impact Statement

Our paper proposes new methods to explain the predictions of decoder-only sequence classification model through input attributions. By providing better attributions, our methods improve the interpretability of ML models, enabling human reviewers to better understand model predictions. Thus, by providing high-quality explanations, our work improves the trustworthiness of ML models, supporting the safe and responsible deployment of AI.

Acknowledgements

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co and its affiliates ("J.P. Morgan") and is not a product of the Research Department of J.P. Morgan. J.P. Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020a.

Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385.

Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.-R., and Wolf, L. Xai for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning*, pp. 435–451. PMLR, 2022a.

Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.-R., and Wolf, L. Xai for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning*, pp. 435–451. PMLR, 2022b.

Anders, C. J., Weber, L., Neumann, D., Samek, W., Müller, K.-R., and Lapuschkin, S. Finding and removing clever hans: using explanation methods to debug and improve deep models. *Information Fusion*, 77:261–295, 2022.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. A diagnostic study of explainability techniques for text classification. *arXiv preprint arXiv:2009.13295*, 2020.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chefer, H., Gur, S., and Wolf, L. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021.

Gekhman, Z., Herzig, J., Aharoni, R., Elkind, C., and Szpektor, I. Trueteacher: Learning factual consistency evaluation with large language models. *arXiv preprint arXiv:2305.11171*, 2023.

Hollenstein, N. and Beinborn, L. Relative importance in sentence processing. *arXiv preprint arXiv:2106.03471*, 2021.

Hsiao, C.-R. and Raghavan, T. Shapley value for multichoice cooperative games, i. *Games and economic behavior*, 5(2):240–256, 1993.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kheiri, K. and Karimi, H. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv preprint arXiv:2307.10234*, 2023.

Krishna, S., Ma, J., Slack, D. Z., Ghandeharioun, A., Singh, S., and Lakkaraju, H. Post hoc explanations of language models can improve language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=3H37XciUEv.

Li, Z., Li, X., Liu, Y., Xie, H., Li, J., Wang, F.-l., Li, Q., and Zhong, X. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*, 2023.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pp. 1–17, 2018.

OpenAI. Fine-tuning - openai api. https://platform.openai.com/docs/guides/fine-tuning/use-a-fine-tuned-model, 2023. Accessed: 2024-01-29.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

Rosenthal, S., Farra, N., and Nakov, P. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pp. 502–518, 2017.

Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K. T., Müller, K.-R., and Montavon, G. Higher-order explanations of graph neural networks via relevant walks. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7581–7596, 2021.

Shapley, L. S. et al. A value for n-person games. 1953.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1170.

Strumbelj, E. and Kononenko, I. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017a.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017b.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, N., Yang, H., and Wang, C. D. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*, 2023.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023a.

Wu, Z., Zhang, L., Cao, C., Yu, X., Dai, H., Ma, C., Liu, Z., Zhao, L., Li, G., Liu, W., et al. Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology nli task. *arXiv preprint arXiv:2304.09138*, 2023b.

Young, H. P. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14:65–72, 1985.

Zacharias, J., von Zahn, M., Chen, J., and Hinz, O. Designing a feature selection method based on explainable artificial intelligence. *Electronic Markets*, 32(4):2159–2184, 2022.

Zhang, X., Zhao, J. J., and LeCun, Y. Character-level convolutional networks for text classification. In *NIPS*, 2015.

Zhang, Y., Jiang, M., Meng, Y., Zhang, Y., and Han, J. Pieclass: Weakly-supervised text classification with prompting and noise-robust iterative ensemble training. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12655–12670, 2023.

A. Proofs

Proposition 1. SPPI's attribution $\phi_i = p_i^c - p_{i-1}^c$ satisfies the *local accuracy* property: $p_n^c - p_0^c = \sum_{i=1}^n \phi_i$.

Proof. From Eqn. 6, we have $\phi_i = p_i^c - p_{i-1}^c$. Expanding $\sum_{i=1}^n \phi_i$, we have

$$\sum_{i=1}^{n} \phi_i = \sum_{i=1}^{n} p_i^c - p_{i-1}^c \tag{18}$$

$$\sum_{i=1}^{n} \phi_i = (p_1^c - p_0^c) + (p_2^c - p_1^c) + (p_3^c - p_2^c) + .. \\ ... + (p_n^c - p_{n-1}^c). \tag{19}$$

All the terms except p_0^c and p_n^c cancel out, yielding $\sum_{i=1}^n \phi_i = p_n^c - p_0^c$ □

Proposition 2. Let $\vec{P}^* = [P_1^*, P_2^*, ..., P_{n-1}^*]$ denote the vector representation of the Shapley distribution, where P_i^* denotes the probability of sampling a coalition of size i. Let P^* denote the matrix representation of the Shapley distribution consisting of entries P_{ij}^* that indicate the probability of sampling coalitions of size i, where j is the last active feature. Then, P_{ij}^* can be expressed in terms of P_i^* as follows:

$$P_{ij}^* = \begin{cases} P_i^* \binom{j-1}{i-1} / \binom{n}{i} & \text{if } j \geq i \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Proof. Since we have n features, the total number of ways in which a subset of i can be formed is $\binom{n}{i}$. If $j \geq i$, the total number of subsets where j is the last active feature is given by $\binom{j-1}{i-1}$. Thus, the probability of selecting a subset of i features, where j is the last active feature is $\binom{j-1}{i-1}/\binom{n}{i}$. Multiplying this with the probability of sampling a coalition of size i, we have, $P_{ij}^* = P_i^* \binom{j-1}{i-1}/\binom{n}{i}$ if $j \geq i$. Note that no coalition of size i can be selected such that the index of the last active feature is less than i. Thus, $P_{ij}^* = 0$ when $j < i$. □

Proposition 3. In PI, predictions on an input coalition S'_{ij} (representing the input $\vec{x}'$ in Fig. 4b), yields additional predictions for coalitions of the form $\{S_{kl}^\dagger\}_{k=1}^i$ i.e. coalitions of sizes $1, 2, .., i$ (represented by $D_r^\dagger$ in Fig. 4e). We can view the distribution of these additional coalitions $S_{kl}^\dagger$ as being conditioned on S'_{ij}. Assuming $i, j, k, l \in N$, this conditional distribution is given by

$$P_{kl|ij}^\dagger = \begin{cases} \binom{l-1}{k-1}\binom{j-l}{i-k} / (\binom{j-1}{i-1} i) & \text{if } k \leq i, l \leq j, j \geq i \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

Proof. The total number of coalitions of the form S'_{ij} is given by $\binom{j-1}{i-1}$. For S'_{ij} to yield an intermediate coalition of the form $S_{kl}^\dagger$, we need two conditions to hold:

- Feature l to be the k^{th} active feature.
- There need to be exactly $i - k$ active features after feature l.

There are $\binom{l-1}{k-1}$ possible ways of satisfying the first condition and $\binom{j-l}{i-k}$ ways of satisfying the second. Thus, totally, there are a total of $\binom{l-1}{k-1}\binom{j-l}{i-k}$ possible coalitions of the form S'_{ij} that satisfy both conditions. Expressed as a fraction of the total number of possible coalitions of the form S'_{ij}, this yields $\binom{l-1}{k-1}\binom{j-l}{i-k}/\binom{j-1}{i-1}$. Since we get a total of i intermediate coalitions, we divide by i to obtain the normalized conditional probability $P_{kl|ij}^\dagger = \binom{l-1}{k-1}\binom{j-l}{i-k}/(\binom{j-1}{i-1}i)$. Note that this probability only holds when $k \leq i, l \leq j$ and $j \geq i$. Under all other conditions, there are no intermediate coalitions that satisfy the conditions above, resulting in $P_{kl|ij}^\dagger = 0$.

□

B. Additional Experimental Details

B.1. Training Setup

We train all models for 10 epochs with a learning rate of 5×10^{-5}. We use the Adam optimizer and a batch size of 16. We truncate the inputs when necessary so that it fits within the support input lengths for GPT-2 and Llama-2. For the TrueTeacher dataset, we use the following format in the input: "[Assertion]: *hypothesis* [Document]: *premise*". Note that putting the *hypothesis* up front allows us to make intermediate predictions on masked versions of the premise. For LoRA, we use a rank=16, alpha=32 and lora_dropout=0.1

B.2. Note on LRP and inp x grad

For the LRP and inp x grad methods, we found that taking the l2 norm improves the AUC for both activation and inverse activation studies. This empirical finding is consistent with the results reported in prior work (Atanasova et al., 2020). Thus, to have the best performing version of prior work, we use the l2 norm for computing attributions with LPR and inp x grad.

C. Quantifying MP-PI's Sensitivity to P'

A key component of our proposed MP-PI method is finding an optimal P' that results in the distribution of intermediate samples resembling the Shapley distribution. Table 4 quantifies the marginal benefit of choosing this optimal P' over an alternative sampling scheme of directly using the Shapley distribution to sample. For most datasets we find that the optimized sampling scheme provides higher quality explanations, measured in terms of the AUC of the activation and inverse activation studies.

Table 4. Comparing the performance of MP-PI with the optimal sampling scheme and the default Shapley sampling

Dataset	AUC Act. (↑)		AUC Inv. Act. (↓)	
	Opt	Shap	Opt	Shap
IMDB	**0.9697**	0.9696	**0.6128**	0.6263
SST-2	**0.9562**	0.9461	**0.4261**	0.4307
AG-News	0.8669	**0.8758**	**0.5957**	0.5958
Twitter-Fin	0.847	**0.856**	**0.67**	0.6842
Twitter-Sentiment	**0.8866**	0.8843	**0.2733**	0.2779
Twitter-Emotion	**0.932**	0.9292	**0.2144**	0.2242
TrueTeacher NLI	**0.921**	0.9198	**0.3545**	0.3702

D. Additional Results

D.1. Qualitative Evaluation of Attributions

Fig. 6, 7 compares the attributions produced by different XAI techniques on two mispredicted examples from the IMDB dataset. Note that the example contains a negative movie review, which is mispredicted as a positive review by the model in both cases. We compute the attribution with respect to the predicted class (i.e. the positive class). Attributions are computed at the sentence level. Sentences that support the prediction (i.e. positive sentences) are highlighted in green and ones that don't support the prediction (i.e. negative sentences) are highlighted in red. The shade of red/green indicates the magnitude of the normalized attribution. Note that for a human reviewer to catch this mistake, it is important for the XAI technique to highlight sentences that don't support the prediction. We see that in both cases, A-last and GAE fail to highlight any sentence in red. The attributions provided by inp x grad, LRP and integrated gradients are incorrect as they fail to properly highlight positive and negative sentences[3]. Only Kernel SHAP, SP-PI and MP-PI provide attributions that are consistent with the sentiment of each sentence. This shows that perturbation based attribution methods such as the ones proposed in this paper provide attributions that are the most helpful in the event of a misprediction. Our quantitative results in Section 4.2 support these qualitative findings.

D.2. Plots for Activation and Inverse Activation Studies

Fig. 8, 9 show the plots for activation and inverse activation studies.

D.3. Statistical Significance of Activation and Inverse Activation Studies

Table 5 and Table 6 list the 95% confidence intervals for the mean AUC reported in Table 2 and Table 3 respectively. Note that the width of the confidence intervals is smaller than the magnitude of improvements offered by our proposal over prior works.

D.4. Similarity with SHAP values

Our work uses intermediate predictions to compute input attributions that approximate SHAP values. To validate this claim, we compare the attributions produced by our method with those obtained by running Kernel SHAP with a very high sample budget (budget $=16n$, where n represents the number of features). We plot the distribution of cosine similarity between the attributions to understand how closely the two attributions match up. The results are shown in Fig. 5. We find that there is a high degree of similarity between the attributions provided by MP-PI and that of Kernel SHAP for most datasets.

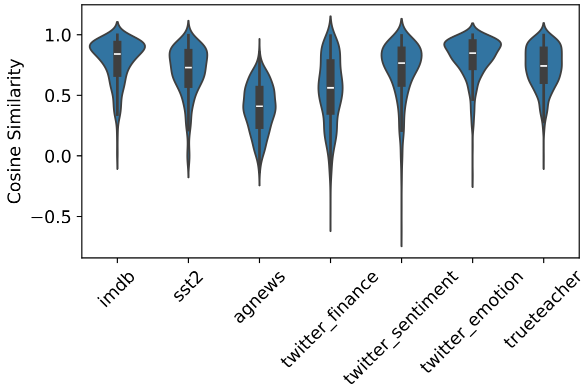


Figure 5. Distribution of cosine similarities between Kernel SHAP and MP-PI attributions. For most datasets, we see a high cosine similarity, indicating that the attributions produced by MP-PI indeed approximates SHAP values.

E. Limitations

While our methods are capable of providing high quality attributions, there are some limitations that need to be considered when using them in practice.

- *Validity of masked inputs:* Our methods compute attributions by considering the prediction of the model on masked/perturbed versions of the input. This assumes that masked versions of the input are valid inputs to the model.
- *Difference with SHAP values:* The attributions computed by MP-PI differ from SHAP values due to two key reasons.

[3] Just for collecting these qualitative samples, we don't take the l2 norm of the attributions for inp x grad and LRP as taking the norm would result in only positive attributions.

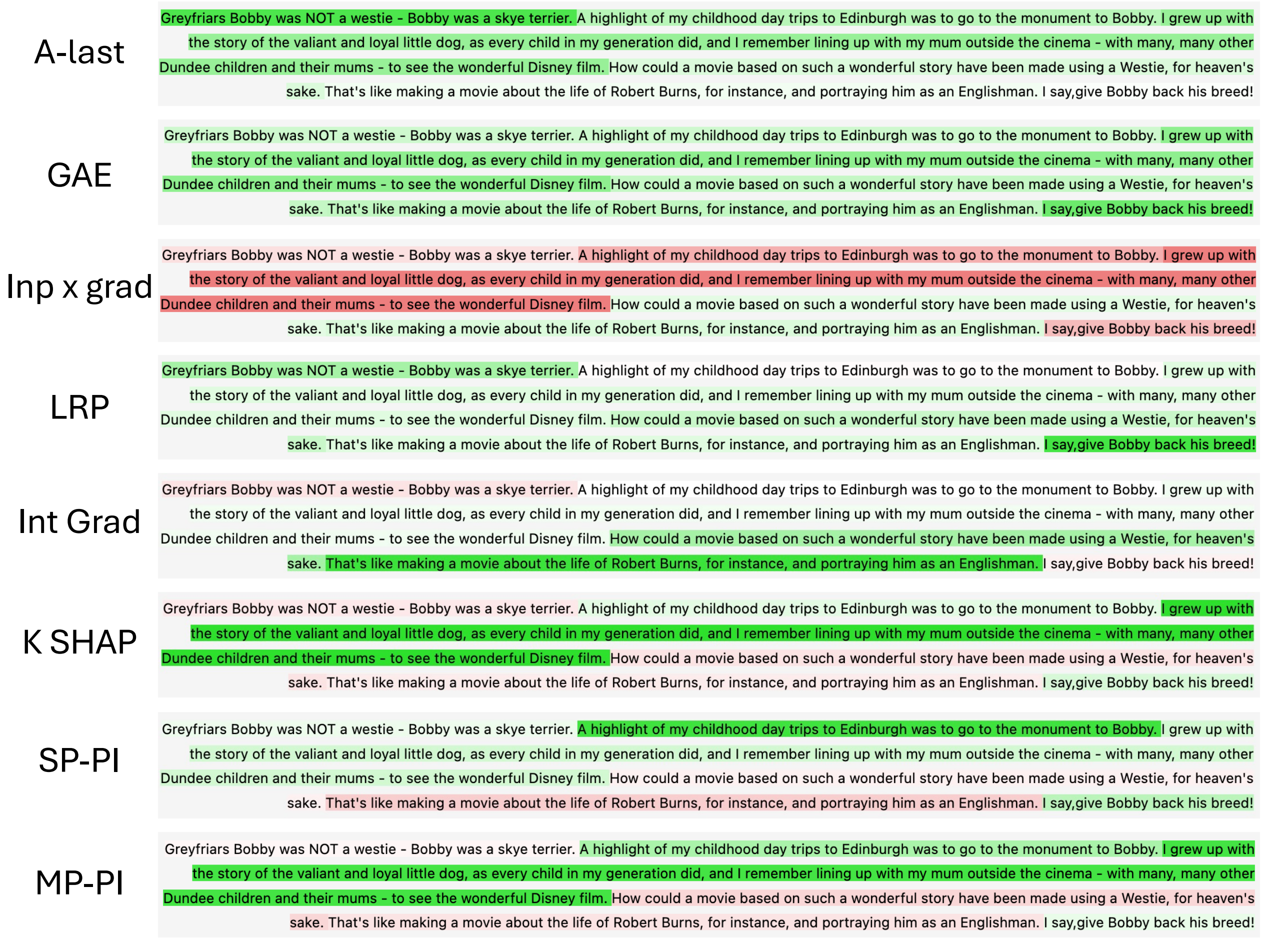
Table 5. 95% Confidence intervals for Activation Study (Table 2)

Method	IMDB	SST-2	AG-news	Twitter-Fin	Twitter-Sen	Twitter-Emo	TrueTeacher
Random	±0.018	±0.024	±0.022	±0.029	±0.03	±0.028	±0.025
A-Last	±0.017	±0.024	±0.022	±0.026	±0.03	±0.027	±0.025
SP-PI	±0.008	±0.019	±0.021	±0.026	±0.027	±0.021	±0.019
GAE	±0.017	±0.021	±0.022	±0.026	±0.029	±0.027	±0.024
Inp x Grad	±0.018	±0.023	±0.022	±0.025	±0.032	±0.028	±0.025
LRP for Transformers	±0.018	±0.023	±0.022	±0.025	±0.032	±0.029	±0.025
Int. Grad	±0.016	±0.019	±0.023	±0.027	±0.029	±0.024	±0.024
Kernel SHAP	±0.009	±0.016	±0.021	±0.024	±0.025	±0.021	±0.022
MP-PI	±0.007	±0.011	±0.019	±0.022	±0.018	±0.015	±0.015

Table 6. 95% Confidence intervals for Inverse Activation Study (Table 3)

Method	IMDB	SST-2	AG-news	Twitter-Fin	Twitter-Sen	Twitter-Emo	TrueTeacher
Random	±0.017	±0.024	±0.023	±0.029	±0.03	±0.027	±0.026
A-Last	±0.02	±0.024	±0.021	±0.03	±0.029	±0.027	±0.029
SP-PI	±0.026	±0.029	±0.024	±0.033	±0.031	±0.028	±0.032
GAE	±0.018	±0.026	±0.022	±0.031	±0.031	±0.026	±0.033
Inp x Grad	±0.017	±0.024	±0.021	±0.03	±0.029	±0.025	±0.034
LRP for Transformers	±0.018	±0.024	±0.021	±0.03	±0.03	±0.024	±0.034
Int. Grad	±0.02	±0.029	±0.023	±0.03	±0.033	±0.029	±0.032
Kernel SHAP	±0.026	±0.03	±0.023	±0.033	±0.03	±0.028	±0.031
MP-PI	±0.026	±0.029	±0.024	±0.035	±0.029	±0.021	±0.031

First, the samples obtained in PI are correlated due to the masked attention mechanism and second, the intermediate predictions may not accurately reflect the prediction of the model on the equivalent masked input.



A-last

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

GAE

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

Inp x grad

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

LRP

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

Int Grad

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

K SHAP

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

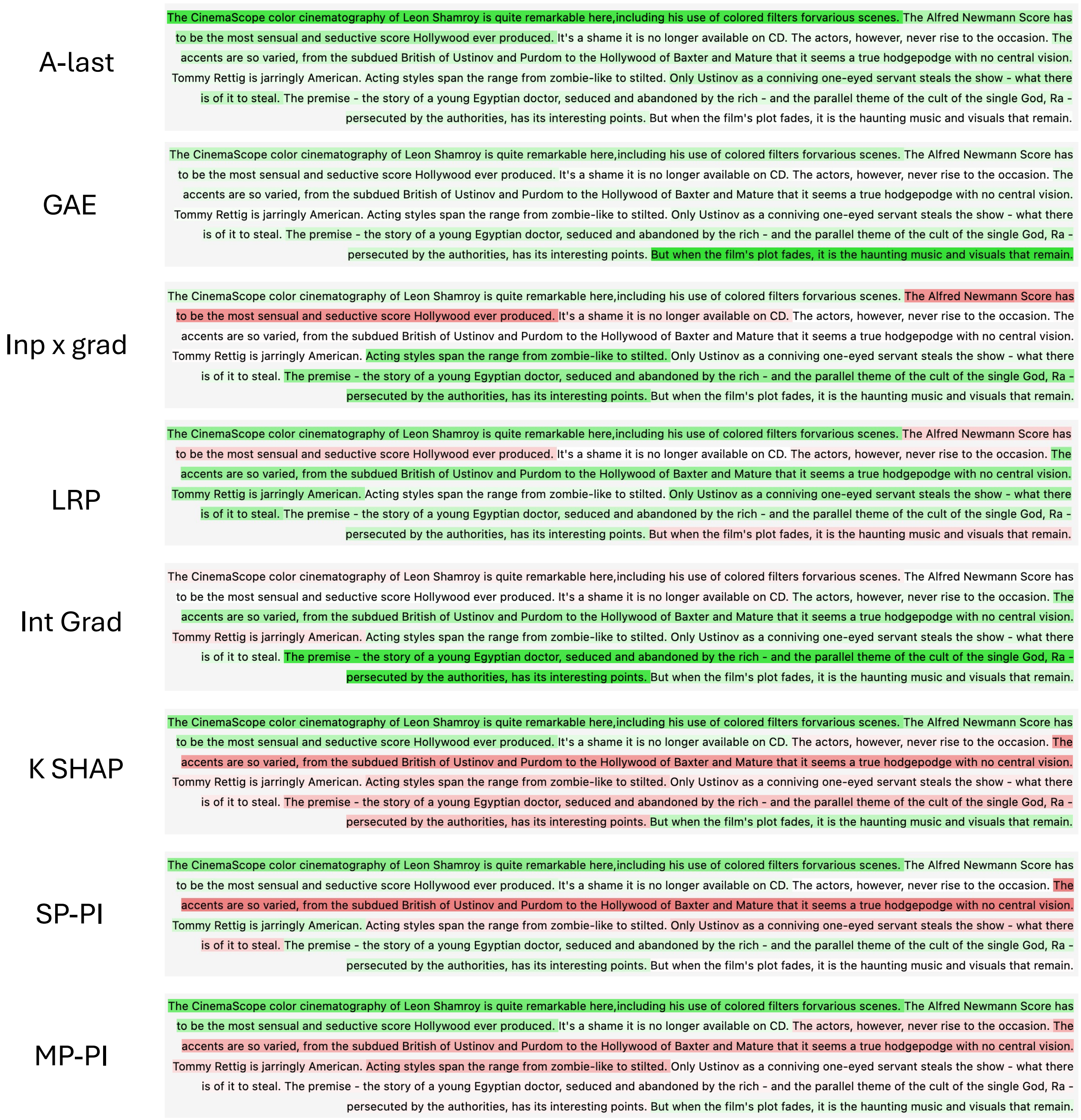
SP-PI

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

MP-PI

Greyfriars Bobby was NOT a westie - Bobby was a skye terrier. A highlight of my childhood day trips to Edinburgh was to go to the monument to Bobby. I grew up with the story of the valiant and loyal little dog, as every child in my generation did, and I remember lining up with my mum outside the cinema - with many, many other Dundee children and their mums - to see the wonderful Disney film. How could a movie based on such a wonderful story have been made using a Westie, for heaven's sake. That's like making a movie about the life of Robert Burns, for instance, and portraying him as an Englishman. I say,give Bobby back his breed!

Figure 6. Comparing the attributions provided by different XAI techniques on a mispredicted sample from the IMDB dataset. The above example is a negative movie review (class 0). The model incorrectly classifies this example as a positive review (class 1). Sentences with positive attributions are highlighted in green and sentences with negative attributions in red. Our proposed methods (SP-PI, MP-PI) provide attributions that are consistent with the sentiment of each sentence, while most prior works provide inconsistent attributions. Faithful explanations such as the ones provided by SP-PI and MP-PI allow a human reviewer to easily identify the misprediction.



The CinemaScope color cinematography of Leon Shamroy is quite remarkable here,including his use of colored filters forvarious scenes. The Alfred Newmann Score has to be the most sensual and seductive score Hollywood ever produced. It's a shame it is no longer available on CD. The actors, however, never rise to the occasion. The accents are so varied, from the subdued British of Ustinov and Purdom to the Hollywood of Baxter and Mature that it seems a true hodgepodge with no central vision. Tommy Rettig is jarringly American. Acting styles span the range from zombie-like to stilted. Only Ustinov as a conniving one-eyed servant steals the show - what there is of it to steal. The premise - the story of a young Egyptian doctor, seduced and abandoned by the rich - and the parallel theme of the cult of the single God, Ra - persecuted by the authorities, has its interesting points. But when the film's plot fades, it is the haunting music and visuals that remain.

The CinemaScope color cinematography of Leon Shamroy is quite remarkable here,including his use of colored filters forvarious scenes. The Alfred Newmann Score has to be the most sensual and seductive score Hollywood ever produced. It's a shame it is no longer available on CD. The actors, however, never rise to the occasion. The accents are so varied, from the subdued British of Ustinov and Purdom to the Hollywood of Baxter and Mature that it seems a true hodgepodge with no central vision. Tommy Rettig is jarringly American. Acting styles span the range from zombie-like to stilted. Only Ustinov as a conniving one-eyed servant steals the show - what there is of it to steal. The premise - the story of a young Egyptian doctor, seduced and abandoned by the rich - and the parallel theme of the cult of the single God, Ra - persecuted by the authorities, has its interesting points. But when the film's plot fades, it is the haunting music and visuals that remain.

The CinemaScope color cinematography of Leon Shamroy is quite remarkable here,including his use of colored filters forvarious scenes. The Alfred Newmann Score has to be the most sensual and seductive score Hollywood ever produced. It's a shame it is no longer available on CD. The actors, however, never rise to the occasion. The accents are so varied, from the subdued British of Ustinov and Purdom to the Hollywood of Baxter and Mature that it seems a true hodgepodge with no central vision. Tommy Rettig is jarringly American. Acting styles span the range from zombie-like to stilted. Only Ustinov as a conniving one-eyed servant steals the show - what there is of it to steal. The premise - the story of a young Egyptian doctor, seduced and abandoned by the rich - and the parallel theme of the cult of the single God, Ra - persecuted by the authorities, has its interesting points. But when the film's plot fades, it is the haunting music and visuals that remain.

The CinemaScope color cinematography of Leon Shamroy is quite remarkable here,including his use of colored filters forvarious scenes. The Alfred Newmann Score has to be the most sensual and seductive score Hollywood ever produced. It's a shame it is no longer available on CD. The actors, however, never rise to the occasion. The accents are so varied, from the subdued British of Ustinov and Purdom to the Hollywood of Baxter and Mature that it seems a true hodgepodge with no central vision. Tommy Rettig is jarringly American. Acting styles span the range from zombie-like to stilted. Only Ustinov as a conniving one-eyed servant steals the show - what there is of it to steal. The premise - the story of a young Egyptian doctor, seduced and abandoned by the rich - and the parallel theme of the cult of the single God, Ra - persecuted by the authorities, has its interesting points. But when the film's plot fades, it is the haunting music and visuals that remain.

The CinemaScope color cinematography of Leon Shamroy is quite remarkable here,including his use of colored filters forvarious scenes. The Alfred Newmann Score has to be the most sensual and seductive score Hollywood ever produced. It's a shame it is no longer available on CD. The actors, however, never rise to the occasion. The accents are so varied, from the subdued British of Ustinov and Purdom to the Hollywood of Baxter and Mature that it seems a true hodgepodge with no central vision. Tommy Rettig is jarringly American. Acting styles span the range from zombie-like to stilted. Only Ustinov as a conniving one-eyed servant steals the show - what there is of it to steal. The premise - the story of a young Egyptian doctor, seduced and abandoned by the rich - and the parallel theme of the cult of the single God, Ra - persecuted by the authorities, has its interesting points. But when the film's plot fades, it is the haunting music and visuals that remain.

The CinemaScope color cinematography of Leon Shamroy is quite remarkable here,including his use of colored filters forvarious scenes. The Alfred Newmann Score has to be the most sensual and seductive score Hollywood ever produced. It's a shame it is no longer available on CD. The actors, however, never rise to the occasion. The accents are so varied, from the subdued British of Ustinov and Purdom to the Hollywood of Baxter and Mature that it seems a true hodgepodge with no central vision. Tommy Rettig is jarringly American. Acting styles span the range from zombie-like to stilted. Only Ustinov as a conniving one-eyed servant steals the show - what there is of it to steal. The premise - the story of a young Egyptian doctor, seduced and abandoned by the rich - and the parallel theme of the cult of the single God, Ra - persecuted by the authorities, has its interesting points. But when the film's plot fades, it is the haunting music and visuals that remain.

The CinemaScope color cinematography of Leon Shamroy is quite remarkable here,including his use of colored filters forvarious scenes. The Alfred Newmann Score has to be the most sensual and seductive score Hollywood ever produced. It's a shame it is no longer available on CD. The actors, however, never rise to the occasion. The accents are so varied, from the subdued British of Ustinov and Purdom to the Hollywood of Baxter and Mature that it seems a true hodgepodge with no central vision. Tommy Rettig is jarringly American. Acting styles span the range from zombie-like to stilted. Only Ustinov as a conniving one-eyed servant steals the show - what there is of it to steal. The premise - the story of a young Egyptian doctor, seduced and abandoned by the rich - and the parallel theme of the cult of the single God, Ra - persecuted by the authorities, has its interesting points. But when the film's plot fades, it is the haunting music and visuals that remain.

The CinemaScope color cinematography of Leon Shamroy is quite remarkable here,including his use of colored filters forvarious scenes. The Alfred Newmann Score has to be the most sensual and seductive score Hollywood ever produced. It's a shame it is no longer available on CD. The actors, however, never rise to the occasion. The accents are so varied, from the subdued British of Ustinov and Purdom to the Hollywood of Baxter and Mature that it seems a true hodgepodge with no central vision. Tommy Rettig is jarringly American. Acting styles span the range from zombie-like to stilted. Only Ustinov as a conniving one-eyed servant steals the show - what there is of it to steal. The premise - the story of a young Egyptian doctor, seduced and abandoned by the rich - and the parallel theme of the cult of the single God, Ra - persecuted by the authorities, has its interesting points. But when the film's plot fades, it is the haunting music and visuals that remain.

Figure 7. Comparing the attributions provided by different XAI techniques on a mispredicted sample from the IMDB dataset. The above example is a negative movie review (class 0). The model incorrectly classifies this example as a positive review (class 1). Sentences with positive attributions are highlighted in green and sentences with negative attributions in red. Our proposed methods (SP-PI, MP-PI) provide attributions that are consistent with the sentiment of each sentence, while most prior works provide inconsistent attributions. Faithful explanations such as the ones provided by SP-PI and MP-PI allow a human reviewer to easily identify the misprediction.

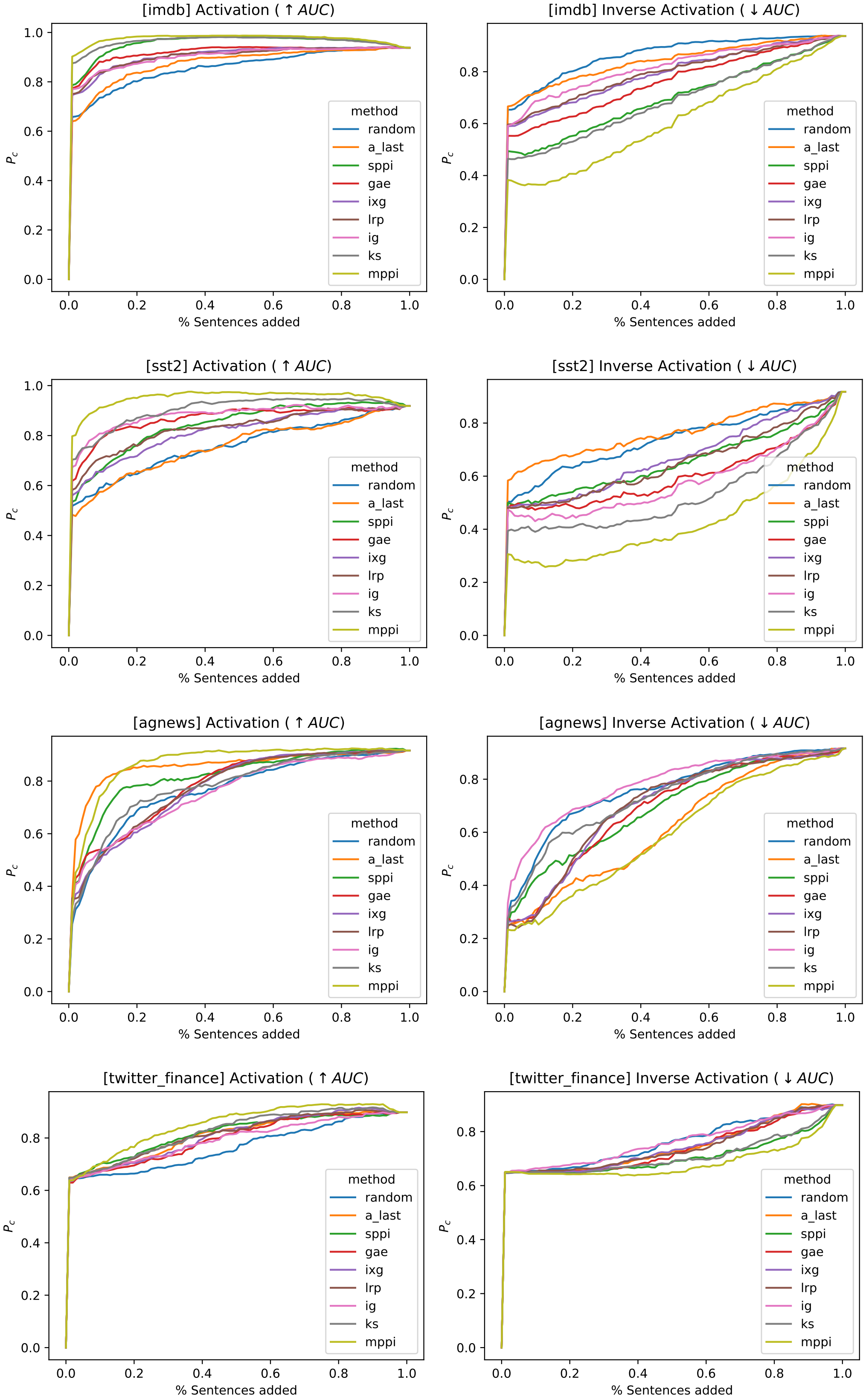


Figure 8. Plots for activation and inverse activation studies.

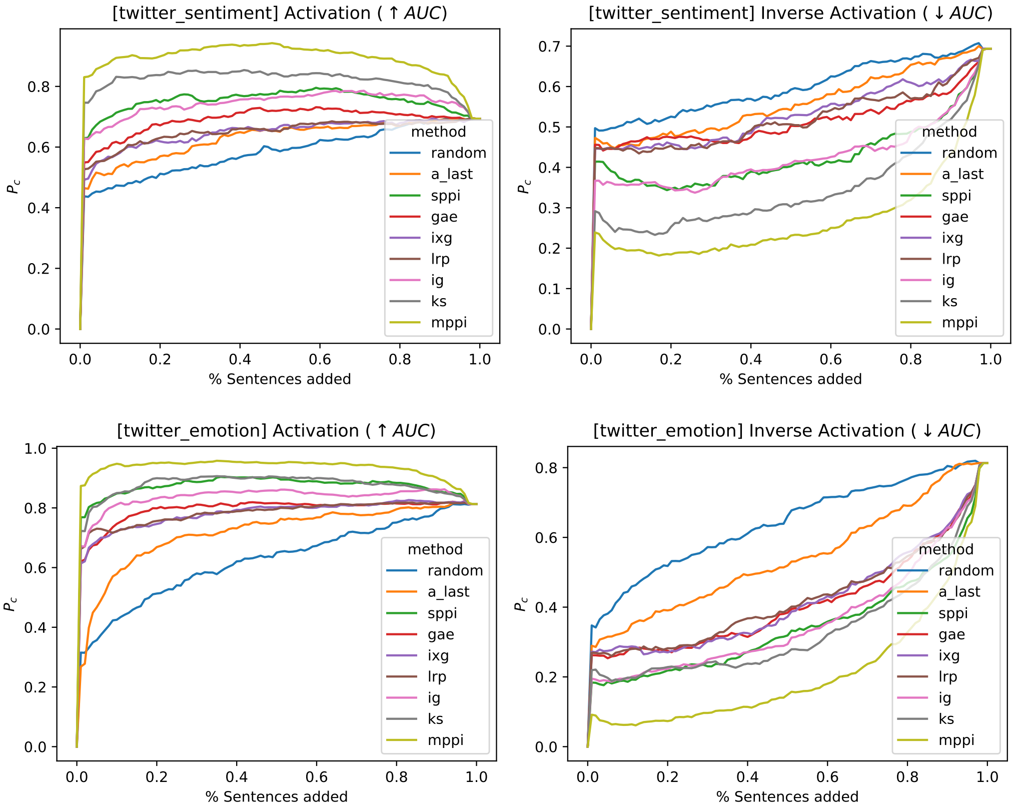


Figure 9. Plots for activation and inverse activation studies.