B-cos LM: Efficiently Transforming Pre-trained Language Models for Improved Explainability

Anonymous authors
Paper under double-blind review

Abstract

Post-hoc explanation methods for black-box models often struggle with faithfulness and human interpretability due to the lack of explainability in current neural architectures. Meanwhile, B-cos networks have been introduced to improve model explainability by proposing an architecture that removes bias terms and promotes input-weight alignment. Although B-cos networks have shown success in building explainable systems, their application has so far been limited to computer vision models and their associated training pipelines. In this work, we introduce B-cos LMs, i.e., B-cos language models (LMs) empowered for natural language processing (NLP) tasks. Our approach directly transforms pre-trained language models into B-cos LMs by combining B-cos conversion and task fine-tuning, improving efficiency compared to previous methods. Our automatic and human evaluation results demonstrate that B-cos LMs produce more faithful and human interpretable explanations than post-hoc methods, while maintaining task performance comparable to conventional fine-tuning. Our in-depth analysis explores how B-cos LMs differ from conventionally fine-tuned models in their learning processes and explanation patterns. Finally, we present a first exploration of transforming decoder-only models to B-cos LMs for generation tasks.

1 Introduction

Pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2019; Brown et al., 2020; OpenAI, 2023) have significantly advanced performance across a plethora of NLP tasks (Wang et al., 2018; Gao et al., 2023). However, their complex architectures and black-box nature make understanding their behavior a persistent challenge (Bommasani et al., 2021). To address this, research has increasingly focused on understanding model predictions in various natural language understanding and generation tasks using different forms of explanations, such as input-based explanations (Feng et al., 2024; Wei Jie et al., 2024; Jiang et al., 2024; Madsen et al., 2024; Yin & Neubig, 2022; Deiseroth et al., 2023), natural language explanations (Ramnath et al., 2024; Wang et al., 2025), and concept-based explanations (Yu et al., 2024; Raman et al., 2024). Among others, input-based explanations, often referred to as rationales, aim to reveal how specific inputs influence a model's prediction (Arras et al., 2019; Atanasova et al., 2020; Lyu et al., 2024). In this work, we focus on input-based explanations, as they offer the most direct insight into model behavior and are often mandated by laws, such as the EU Artificial Intelligence Act.

Most input-based explanation methods for neural models are post-hoc, meaning that they attempt to explain a model's behavior only after it has been trained and deployed (Sundararajan et al., 2017; Ribeiro et al., 2016). While these methods are widely used and easy to apply, they have been shown to produce unfaithful explanations that do not accurately reflect the model's actual reasoning process (Kindermans et al., 2019; Slack et al., 2020; Pruthi et al., 2020). They also struggle with human interpretability, making it difficult for users to understand the model's reasoning (Smilkov et al., 2017; Ismail et al., 2021). Prior research suggests that these limitations stem from a lack of inherent explainability in current models, that is, the model's ability to generate faithful and interpretable explanations by design (Kindermans et al., 2018; Alvarez Melis & Jaakkola, 2018; Rudin, 2019). As a result, improving model explainability is crucial for producing explanations that are both reliable and useful to users.[1] Figure 1 provides examples illustrating this issue.

[1] Considering the evolving definition of these terms in past literature, we provide a detailed definition in Appendix A.

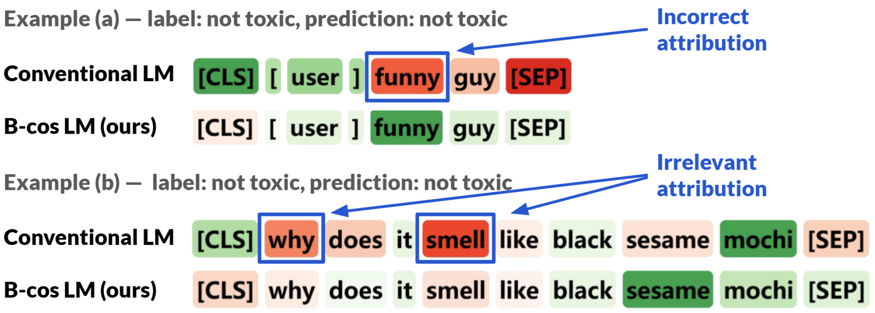

Example (a) — label: not toxic, prediction: not toxic
Incorrect attribution
Conventional LM
[CLS] [user] funny guy [SEP]
B-cos LM (ours)
[CLS] [user] funny guy [SEP]
Example (b) — label: not toxic, prediction: not toxic
Irrelevant attribution
Conventional LM
[CLS] why does it smell like black sesame mochi [SEP]
B-cos LM (ours)
[CLS] why does it smell like black sesame mochi [SEP]

Figure 1: **Visualization of W(x)x in a conventionally fine-tuned model (Conventional LM) and a B-cos LM.** Green (red) indicates the positive (negative) impact of tokens on the prediction. In both examples, both models correctly predict *not toxic*. In the Conventional LM, "funny" is incorrectly assigned a negative attribution in example (a), and in example (b), irrelevant words like "why" and "smell" are highlighted, making the explanations unfaithful and less interpretable. Examples and explanations are drawn from HateXplain. See §3 for details on how $\mathbf{W}(\mathbf{x})\mathbf{x}$ is computed.

To overcome these limitations, we introduce **B-cos LM**, a dynamic linear model that learns the most task-relevant patterns through increased input-weight alignment pressure. Building upon B-cos networks that were first introduced by Böhle et al. (2022) for computer vision, we ensure the explainability of B-cos LMs through mathematically grounded architectural and computational adaptations, with specialized architectural modifications and training pipelines tailored for NLP tasks.

We conduct comprehensive empirical experiments using encoder-only models on classification tasks. Our focus on classification is motivated by its prevalence in high-stakes applications, such as loan approvals, hiring decisions, and hate speech detection, where explainability is crucial and often legally mandated. Encoder-only models have also seen renewed interest in the research community (Warner et al., 2024; Breton et al., 2025; Chaffin, 2025), and they remain the standard architecture for text classification and continue to perform competitively compared to large language models (LLMs) (Zhao et al., 2024). Beyond that, we also explore applying B-cos LMs to decoder-only models for generation tasks and show that B-cos LMs can be extended to a variety of tasks and the latest model architectures. Our contributions are as follows:

1. We propose B-cos LM, a novel model with enhanced explainability. Automatic and human evaluations demonstrate that B-cos LMs generate more faithful and human interpretable explanations than post-hoc explanations while maintaining a strong task performance.
2. We investigate different strategies for transforming PLMs into task-specific B-cos LMs. Our findings show that combining task fine-tuning and B-cos conversion is the most efficient approach, leading to faster convergence than previous B-cos methods and conventional fine-tuning.
3. We thoroughly investigate how B-cos LMs differ from conventionally fine-tuned models and examine how alignment pressure influences their behavior.
4. We are also the first to explore the transformation of decoder-only models to B-cos LMs for generation tasks, providing a step towards a broader application of B-cos LMs in the era of LLMs.

2 Related Work

Post-hoc Explanation Methods Various methods have been proposed to provide post-hoc explanations for neural model predictions (Atanasova et al., 2020). These methods can be broadly categorized based on how they generate explanations: gradient-based (Simonyan et al., 2014; Kindermans et al., 2016; Sundararajan et al., 2017; Enguehard, 2023), propagation-based (Bach et al., 2015; Shrikumar et al., 2017; Springenberg et al., 2015; Ferrando et al., 2023; Modarressi et al., 2022; 2023), and perturbation-based methods (Li et al., 2016; Ribeiro et al., 2016; Lundberg & Lee, 2017; Deiseroth et al., 2023). Besides, the attention mechanism (Bahdanau et al., 2015) is often viewed as an explanation, particularly in transformer-based models (Vaswani et al., 2017). While most existing work focuses on understanding model predictions in classification settings, recent efforts have also aimed to explain model behavior in generation tasks, including sentence completion (Yin & Neubig, 2022; Ferrando et al., 2023), question answering (Enouen et al., 2024), and summarization (Cohen-Wang et al., 2024).

Although post-hoc methods have been widely used, numerous studies have shown that they lack faithfulness, often failing to capture the true decision-making process of the model (Kindermans et al., 2019; Jain & Wallace, 2019; Slack et al., 2020; Pruthi et al., 2020). Furthermore, they are noisy and may select irrelevant information leading to explanations that cannot be interpreted by humans (Smilkov et al., 2017; Ismail et al., 2021).

From Post-hoc Explanations to Explainable Models Prior research suggests that the lack of faithfulness and human interpretability in post-hoc explanations arises from the fundamental lack of explainability in modern neural models, which are typically optimized solely for task performance (Kindermans et al., 2018; Rudin, 2019; Atanasova et al., 2022). In response, various efforts have been made to enhance model explainability. Some works have introduced constraints that improve specific explanation properties, such as faithfulness (Tutek & Šnajder, 2022; Moradi et al., 2020; 2021; Barkan et al., 2024), consistency (Atanasova et al., 2022), locality (Alvarez Melis & Jaakkola, 2018), and plausibility (Ismail et al., 2021). However, as these constraints are typically imposed as regularizers, their effectiveness in improving explanation quality is not guaranteed (Pruthi et al., 2020). Others have proposed self-explanatory model architectures such as rationale-based models that utilize an "explain-then-predict" pipeline, where one module selects rationales for another to make predictions based on them (Lei et al., 2016). Although seemingly transparent, both components rely on neural networks, making the rationale extraction and utilization processes opaque (Zheng et al., 2022; Jacovi & Goldberg, 2021). Besides, such models may face optimization challenges that limit their practicality in real-world tasks (Lyu et al., 2024).

To tackle these shortcomings, Böhle et al. (2022) proposed B-cos networks. Unlike methods that impose external constraints, B-cos networks improve explainability through mathematically grounded architectural and computational adaptations. Moreover, these adaptations are designed as drop-in replacements for conventional model components, making B-cos networks easy to train with minimal performance loss. Most recently, Arya et al. (2024) explored B-cosification techniques to convert existing models into B-cos models, which reduces the training costs of adopting B-cos architectures.

Despite their successful application in vision tasks, B-cos networks have yet to be explored in NLP, where input modalities and training paradigms differ significantly. In this work, we adapt B-cos models for the language domain, integrating them efficiently into NLP pipelines.

3 Methodology

In this section, we outline the architecture and training process of B-cos LMs and how their design ensures faithful and human interpretable explanations. We first introduce B-cos networks (§ 3.1 and § 3.2) and then describe how we transform PLMs to task-specific B-cos LMs (§ 3.3). Finally, we demonstrate how to generate explanations from B-cos LMs (§ 3.4). Notations used in the work are detailed in Appendix B.

3.1 B-cos Networks

Complex neural networks can be interpreted as generalized linear models (Nair & Hinton, 2010; Alvarez Melis & Jaakkola, 2018; Srinivas & Fleuret, 2019). For each input $\mathbf{x}$, the network effectively applies a linear transformation: $\mathbf{f}(\mathbf{x}) = \mathbf{W}(\mathbf{x})\mathbf{x} + \mathbf{b}(\mathbf{x})$, where both the weight $\mathbf{W}(\mathbf{x})$ and bias $\mathbf{b}(\mathbf{x})$ depend on $\mathbf{x}$. Given that many activation functions are (approximately) piecewise linear, the overall network can be viewed as (approximately) piecewise affine (Alvarez Melis & Jaakkola, 2018). Earlier work refers to such models as dynamic linear models (Böhle et al., 2021; 2022), highlighting the fact that the weight and bias terms dynamically change according to $\mathbf{x}$.

Under this dynamic linear perspective, the linear mapping $\mathbf{W}(\mathbf{x})$ can be seen as attributing model predictions to individual input features, and $\mathbf{W}(\mathbf{x_i})\mathbf{x_i}$ can be seen as the contribution of feature $\mathbf{x_i}$ to the model prediction. However, two challenges hinder the direct use of this interpretation. First, $\mathbf{W}(\mathbf{x})$ alone provides an incomplete and unfaithful model summary since $\mathbf{f}(\mathbf{x}) \neq \mathbf{W}(\mathbf{x})\mathbf{x}$ due to the presence of the bias term

$\mathbf{b}(\mathbf{x})$, and incorporating $\mathbf{b}(\mathbf{x})$ into explanations is highly non-trivial (Wang et al., 2019). Second, $\mathbf{W}(\mathbf{x_i})\mathbf{x_i}$ is often difficult for humans to interpret, as $\mathbf{W}(\mathbf{x})$ does not necessarily align only with task-relevant input patterns (Smilkov et al., 2017) and therefore yields noisy and irrelevant explanations. Figure 1 illustrates these challenges. To address these issues, Böhle et al. (2022) introduced B-cos networks by replacing the conventional linear transformation:

$$\mathbf{f}(\mathbf{x};\mathbf{w},\mathrm{b}) = \mathbf{w}^{\mathbf{T}}\mathbf{x} + \mathrm{b} = \|\mathbf{w}\|\|\mathbf{x}\|\cos(\mathbf{x},\mathbf{w}) + \mathrm{b} \tag{1}$$

with a B-cos transformation:

$$\begin{aligned}\text{B-cos}(\mathbf{x};\mathbf{w}) &= \hat{\mathbf{w}}^{\mathbf{T}}\mathbf{x} \times |\cos(\mathbf{x},\hat{\mathbf{w}})|^{\text{B-1}} \\ &= \|\hat{\mathbf{w}}\|\|\mathbf{x}\||\cos(\mathbf{x},\hat{\mathbf{w}})|^{\mathrm{B}} \times \operatorname{sgn}(\cos(\mathbf{x},\hat{\mathbf{w}}))\end{aligned} \tag{2}$$

where $\hat{\mathbf{w}}$ is a scaled version of $\mathbf{w}$ with unit norm and sgn denotes the sign function.

B-cos$(\mathbf{x};\mathbf{w})$ can be seen as a linear transformation of $\mathbf{x}$ with the dynamic linear weight $\mathbf{w}(\mathbf{x}) = |\cos(\mathbf{x},\hat{\mathbf{w}})|^{\text{B-1}} \times \hat{\mathbf{w}}$. The absence of $\mathbf{b}(\mathbf{x})$ ensures the completeness of summary $\mathbf{w}(\mathbf{x})$. We demonstrate that this completeness extends to an entire network composed of bias-free, dynamic linear modules in § 3.4. Moreover, since the B-cos module output is bounded by $\|\mathbf{x}\|$, the weight $\mathbf{w}$ must align closely with task-relevant patterns to achieve a high cosine similarity and strong activation, especially under additional alignment pressure (B>1). This drives the model to assign greater weight to the most relevant features when optimizing target output probabilities, promoting the learning of representative patterns during training. Consequently, during explanation generation, task-relevant features $\mathbf{x_i}$ receive higher attribution $\mathbf{W}(\mathbf{x_i})\mathbf{x_i}$ due to stronger alignment, while irrelevant features receive lower attribution, suppressed by weaker alignment and the exponential scaling. For a more detailed discussion of how the B-cos transformation enhances faithfulness and human interpretability, see Böhle et al. (2022; 2024).

While early B-cos models were trained from scratch, Arya et al. (2024) recently introduced B-cosification, an efficient method to obtain B-cos models. This approach first modifies conventional models with task capacities to adopt the B-cos architecture, followed by fine-tuning on downstream datasets for B-cos conversion. B-cosified models generate explanations as faithful and interpretable as B-cos models trained from scratch but at a much lower training cost. However, directly applying B-cosification to LMs is non-trivial and inefficient due to the significant differences in model architectures and training pipelines.

3.2 Dynamic Linear Representation of Model Components

Here we describe how each model component in transformers can function as or be converted to a bias-free, dynamic linear module in B-cos LMs.

B-cos Layers B-cos layers are designed as bias-free, dynamic linear modules with a dynamic linear weight matrix $\mathbf{W}(\mathbf{x}) = |\cos(\mathbf{x},\hat{\mathbf{W}})|^{\text{B-1}} \otimes \hat{\mathbf{W}}$. Here, $\otimes$ scales the rows of the matrix $\hat{\mathbf{W}}$ to its right by the scalar entries of the vector to its left.

Non-linear Activation Functions In transformer models, non-linearity is typically introduced using (approximately) piecewise linear activation functions, such as ReLU (Nair & Hinton, 2010) and GELU (Hendrycks & Gimpel, 2016). These functions can be easily interpreted as linear transformations with input-dependent weights. For example, $\text{GELU}(\mathbf{x}) = \mathbf{x} \times (0.5 + 0.5 \times \text{erf}(\mathbf{x}/\sqrt{2}))$ can be interpreted as a linear transformation where the second term acts as the dynamic linear weight.

Attention Blocks Böhle et al. (2024) showed that attention computations can be seamlessly integrated into B-cos networks as a dynamic linear module:

$$\text{Att}(\mathbf{X};\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\mathbf{X}^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}\mathbf{K}\mathbf{X})\mathbf{V}\mathbf{X} = \mathbf{A}(\mathbf{X})\mathbf{V}\mathbf{X} = \mathbf{W}(\mathbf{X})\mathbf{X} \tag{3}$$

For multi-head self-attention (MSA), the output can be viewed as the concatenation of the outputs from H attention heads, followed by a linear projection with matrix $\mathbf{U}$:

$$\text{MSA}(\mathbf{X}) = \mathbf{U}[\mathbf{W}_1(\mathbf{X})\mathbf{X}, ..., \mathbf{W}_H(\mathbf{X})\mathbf{X})] \tag{4}$$

Since this operation maintains a dynamic linear structure, the multi-head attention block remains a dynamic linear module.

Normalization Layers Following Böhle et al. (2024), we apply bias-free normalization layers to ensure completeness of explanations:

$$\star\mathbf{Norm}(\mathbf{x}, \mathcal{X}; \gamma) = \frac{\mathbf{x} - \langle \mathcal{X} \rangle_\star}{\sqrt{\mathrm{var}_\star(\mathcal{X})}} \times \gamma \tag{5}$$

where $\mathcal{X}$ represents a batch or sequence of inputs and $\star$ is the dimension along which the mean $\langle\cdot\rangle$ and variance $\mathrm{var}(\cdot)$ are computed (e.g., across the batch or layer). Unlike standard normalization, this variant omits the bias term in the affine transformation to preserve explanation completeness. If a running mean estimate is used during inference, the centering term $\langle \mathcal{X} \rangle_\star$ is also removed. This yields a bias-free, dynamic linear transformation with weight $\sqrt{\mathrm{var}_\star^{-\mathbf{1}}(\mathcal{X})}\times\gamma$.

3.3 B-cosification for LMs

In this section, we present our B-cosification approach for LMs. We summarize the differences between B-cosification for LMs, its counterpart for vision models, and conventional fine-tuning in Table 1.

Property	**Conventional Fine-tuning**	**B-cosification for vision** (Arya et al., 2024)	**B-cos LM (ours)**
Bias terms	yes	no	no
B (alignment pressure)	1	2	1.25 / 1.5
Pred. Head Activations	tanh	n/a[2]	identity
Prior task abilities	no	yes	no
Training objectives	Task fine-tuning	B-cos conversion	Task fine-tuning & B-cos conversion

Table 1: Comparison between conventional fine-tuning, B-cosification for computer vision models and B-cosification for language models (B-cos LM). Conventional fine-tuning and B-cosification for vision follow the configuration of BERT for sequence classification and CLIP (Radford et al., 2021), respectively (cf. § 3 for details).

3.3.1 B-cos Adaptations

Given a conventional model, we first modify its architecture and computation to integrate the B-cos framework.

Architectural Adaptations For completeness and faithfulness of explanations, we follow Arya et al. (2024) and remove all bias terms in models, including those in the affine transformations of layer normalization and attention blocks. Additionally, a prediction head is typically added on top of the transformer before fine-tuning for downstream tasks in the NLP pipeline. This head often includes activation functions that are not (approximately) piecewise linear, such as sigmoid and tanh. To accommodate the unique architecture of LMs, we remove all activation functions in the prediction heads, as they generate explanations that are not locally difference-bounded (Alvarez Melis & Jaakkola, 2018) and introduce numerical instability during explanation generation. Our experiments show that the added non-linearity from B>1 could compensate for this removal.

Introducing B-cos Computation To promote input-weight alignment and improve human interpretability of explanations, we replace all linear transformations with B-cos transformations in § 3.1. For a more efficient B-cosification, B-cos layers are initialized with the corresponding weights of the original model.

3.3.2 Fine-tuning

The B-cos adaptations above modify the architecture and computation of models, requiring fine-tuning to restore their capabilities and adapt to alignment pressure. Following the "pre-train then fine-tune" paradigm,

[2] Arya et al. (2024) used a single linear layer on top of CLIP so the prediction head activation is not applicable in their setup.

which is frequently utilized in NLP tasks, we directly transform PLMs to B-cos LMs, rather than adapting task-specific models as done in previous work (Arya et al., 2024). This fundamental difference in the training pipeline adds complexity to B-cosification for LMs, as the objective involves both B-cos conversion and task fine-tuning. While there are multiple ways to conjoin these two steps (cf. § 5), we find that the most efficient way is to combine them by first applying B-cos adaptations to a PLM and then fine-tuning it on a downstream task. Following Böhle et al. (2022), we use the binary cross-entropy (BCE) loss instead of the conventional cross-entropy loss, as it explicitly maximizes the absolute target logits and strengthens the alignment pressure. We provide an extensive comparison of different B-cosification setups in § 5.

3.4 Computing B-cos Explanations

Once trained, the B-cos LM can generate explanations that faithfully summarize its decision-making process during inference. As all components are dynamic linear with no bias terms (cf. § 3.2), the entire model computation can be expressed as a sequence of matrix multiplications, which can be completely summarized as a single dynamic linear function:

$$\hat{\mathbf{W}}_L(\mathbf{A}_L)\hat{\mathbf{W}}_{L-1}(\mathbf{A}_{L-1})...\hat{\mathbf{W}}_1(\mathbf{A}_1 = \mathbf{X})\mathbf{X} = \Pi_{j=1}^{L}\hat{\mathbf{W}}_j(\mathbf{A}_j) \tag{6}$$

Note that a residual connection of $\mathbf{W}(\mathbf{x})\mathbf{x} + \mathbf{x}$ with $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{W}(\mathbf{x}) \in \mathbb{R}^{n \times n}$ is mathematically equivalent to a single dynamic linear transformation of $(\mathbf{W}(\mathbf{x}) + \mathbf{I}_n)\mathbf{x}$. Considering the textual inputs specific to LMs, we attribute the model's predictions to the embedding representations. Specifically, to quantify the contribution of a token i to a model prediction, we compute the dot product $\mathbf{W}(\mathbf{x}_i)\mathbf{x}_i$ between its embedding $\mathbf{x}_i$ and the corresponding dynamic linear weight $\mathbf{W}(\mathbf{x}_i)$ for the target class logit. For the remainder of the paper, we will refer to such explanations as *B-cos explanations*.

4 Experiments

We evaluate the task performance of B-cos LMs and faithfulness of B-cos explanations with automatic evaluation across various tasks and PLMs. In addition, we conduct a human evaluation study to compare the human interpretability of B-cos explanations. §4.1–4.3 describe our automatic evaluation setup, results, as well as human evaluation study, respectively. §4.4 provides a qualitative analysis. Finally, we conduct an ablation study in §4.5. More details on the experimental setup and baseline methods are provided in Appendix C and a comparison of computational efficiency is provided in Appendix I.1.

4.1 Experimental Setup

Datasets and Models Our experiments use three datasets: AG News (topic classification, Zhang et al., 2015), IMDB (sentiment analysis, Maas et al., 2011), and HateXplain (hate speech detection, Mathew et al., 2021). BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and RoBERTa (Liu et al., 2019) are used as the basis for conventional fine-tuning and for obtaining B-cos LMs. We set B=1.25 for IMDB and B=1.5 for AG News and HateXplain datasets.

Baselines We compare B-cos explanations against a diverse set of post-hoc explanation methods: Attention (Bahdanau et al., 2015), InputXGradient (IxG, Kindermans et al., 2016), Sequential Integrated Gradients (SIG, Enguehard, 2023), DecompX (Modarressi et al., 2023), Shapley Value Sampling (ShapSampl, Strumbelj & Kononenko, 2010), and LIME (Ribeiro et al., 2016). We also apply these methods to a model trained with Saloss (Chrysostomou & Aletras, 2021), which incorporates additional faithfulness regularization. This setup enables a direct comparison between B-cos LMs and models specifically optimized for explainability. For embedding-level explanation methods, we aggregate attributions by summing across all embedding dimensions.

Faithfulness Metrics For a more comprehensive evaluation, we employ two different methods to assess faithfulness. First, we report two perturbation-based metrics (DeYoung et al., 2020):

- **Comprehensiveness** (Comp) measures the average drop in predicted class probability after masking out the top $k\%$ most important tokens in the explanation. A higher score indicates better faithfulness.
- **Sufficiency** (Suff) measures the average drop in predicted class probability after keeping only the top $k\%$ tokens. A lower score indicates better faithfulness.

To avoid arbitrary choices of k, we compute Comp and Suff for multiple values ($k = 10, 20, ..., 90$) and summarize them using the Area Over the Perturbation Curve (AOPC, DeYoung et al., 2020).

In addition, we introduce a new faithfulness metric called Sequence Pointing Game (SeqPG), inspired by the grid pointing game in vision tasks (Böhle et al., 2021):

- **Sequence Pointing Game** (SeqPG). We evaluate models on synthetic sequences composed of segments associated with different classes. To assess faithfulness, we measure the proportion of positive attribution assigned to the corresponding segment of each class and compute their average. A higher score indicates better faithfulness.

Compared to perturbation-based metrics, SeqPG does not rely on perturbations and thus avoids the potential distortions introduced by token masking. When constructing SeqPG examples, we truncate each segment to a fixed length and randomize segment order to control for length and position effects. We generate synthetic examples using correctly and most confidently classified test instances. SeqPG can be seen as a standardized version of hybrid document evaluation (Poerner et al., 2018). We provide an example of SeqPG in Figure 7 and more details in Appendix D.

4.2 Automatic Evaluation Results

Task Performance Figure 2 shows the accuracy of conventionally fine-tuned, Saloss and B-cos BERT across three datasets (we provide results for DistilBERT and RoBERTa in Appendix E). On AG News and HateXplain, B-cos LMs performs on par with conventional models, with only a minor drop (∼1%) in accuracy. They also outperform Saloss models on these datasets. Only for IMDB, we find a slightly larger drop of 3.06% compared to conventional BERT, though the performance remains strong overall.

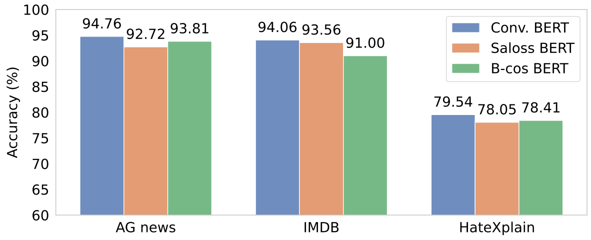


Figure 2: Mean accuracy of conventionally fine-tuned, Saloss and B-cos BERT models averaged over three runs. We use B=1.5, 1.25, and 1.5 for AG News, IMDB, and HateXplain, respectively. B-cos models perform comparably to conventional models on most tasks.

Faithfulness Results Table 2 shows the faithfulness scores for post-hoc explanation methods on conventionally fine-tuned and Saloss BERT models, as well as B-cos explanations from B-cos BERT. The results show that B-cos explanations are consistently and substantially more faithful than post-hoc methods across all models and datasets. On average, B-cos explanations outperform the strongest post-hoc methods on conventional models by 14.63 points in Comp and achieve negative Suff scores, indicating that the identified important tokens alone enable even more confident predictions. B-cos also shows significant gains in SeqPG. While Saloss improves faithfulness for some post-hoc methods over conventional models, it still underperforms compared to B-cos LMs by a large margin. Similar trends are observed for DistilBERT and RoBERTa (Appendix F) as well, further strengthening our findings. Although we do not include rationale-based models in the main experiments because they typically require additional supervision, a supplementary comparison in Appendix G shows that B-cos BERT still outperforms a rationale-based model on HateXplain.

4.3 Human Evaluation

Contrary to previous B-cos studies that rely solely on automatic evaluations to assess explanations, we conduct the first human study to better evaluate the human interpretability and agreement of B-cos explanations. We compare B-cos explanations against three strong post-hoc explanation methods on the conventional BERT model.

Model	Method	AG News			IMDB			HateXplain		
		Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)
Conv. BERT	Attention	24.40	8.09	50	26.84	14.56	50	27.64	13.83	50
	IxG	15.28	10.19	45.41	18.29	16.96	49.42	19.16	18.90	47.24
	SIG	27.02	3.40	64.77	29.34	14.05	59.09	37.31	5.10	66.38
	DecompX	52.16	0.92	84.48	57.94	2.41	63.27	44.86	2.72	66.76
	ShapSampl	43.96	0.46	82.87	58.29	2.44	**71.29**	44.86	2.43	67.17
	LIME	44.95	0.06	80.28	51.45	6.07	60.15	22.64	14.30	57.61
Saloss BERT	Attention	34.73	3.65	50	27.59	13.64	50	34.95	26.26	50
	IxG	14.98	12.66	51.01	24.19	16.30	49.02	26.61	30.94	50.74
	SIG	16.70	8.22	63.74	45.44	8.48	54.96	44.53	21.50	54.70
	DecompX	59.37	0.30	75.34	59.42	5.38	62.02	58.71	13.23	65.17
	ShapSampl	37.73	0.77	73.96	65.38	3.17	70.23	57.05	15.10	72.36
	LIME	53.18	2.37	76.16	53.31	6.32	58.65	21.73	21.96	55.71
B-cos BERT	B-cos	**64.22**	**-1.26**	**87.92**	**74.18**	**-2.87**	70.43	**59.66**	**-4.89**	**77.57**

Table 2: Faithfulness evaluation for conventionally fine-tuned BERT, Saloss BERT and B-cos BERT across three datasets. We use B=1.5, 1.25, and 1.5 for AG News, IMDB and HateXplain, respectively. The best results are in **bold**. We find that B-cos explanations are consistently more faithful than post-hoc explanations from both models.

Following the practice in Enguehard (2023) and Yue et al. (2022), we randomly select 50 instances, respectively, from AG News and HateXplain where the B-cos and conventional models make the same prediction. Five annotators then rate the explanations in terms of human interpretability (how well they understand them) and human agreement (how much they agree with them) on a scale of 1-5. Before the annotation began, the annotators were provided with a clear description of the evaluation task, metrics, and rating scales. They were also shown example annotations along with the reasoning behind each rating to help them better understand the evaluation criteria. Further details on the evaluation criteria, rating scales, annotator instructions, and example annotations can be found in Appendix H.

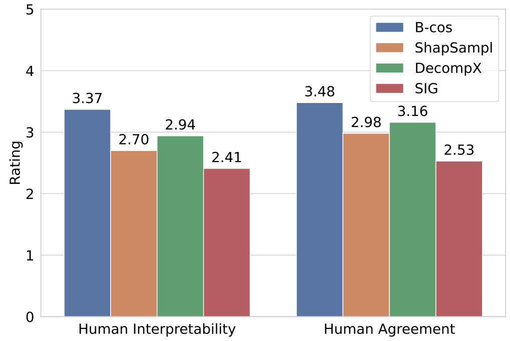


Figure 3: Human evaluation reveals that B-cos explanations have better human interpretability and human agreement than baseline methods.

Figure 3 shows that B-cos explanations have a better human interpretability and exhibit greater alignment with human reasoning than post-hoc methods, even though they are not directly optimized for human agreement. Paired t-tests with a Bonferroni-corrected significance level $\alpha = \frac{0.05}{6} = 0.008\overline{3}$ (Bonferroni, 1936) shows that the improvements of B-cos explanations are statistically significant ($p < \alpha$) for both metrics.

4.4 Qualitative Analysis

Figure 4 provides an example of B-cos and other (post-hoc) explanations. It can be seen that the B-cos explanation highlights important tokens correctly with little focus on irrelevant ones. In contrast, ShapSampl attributes the highest importance to the [SEP] token and provides only little useful information. Meanwhile, DecompX extracts a significant amount of irrelevant information. Overall, we find that the B-cos explanation is more interpretable to humans by providing clearer and more relevant attributions compared to the post-hoc explanations.

4.5 Ablation Study

To better understand B-cos LMs, we conduct an ablation study evaluating the impact of key design choices on task performance and explanation faithfulness.

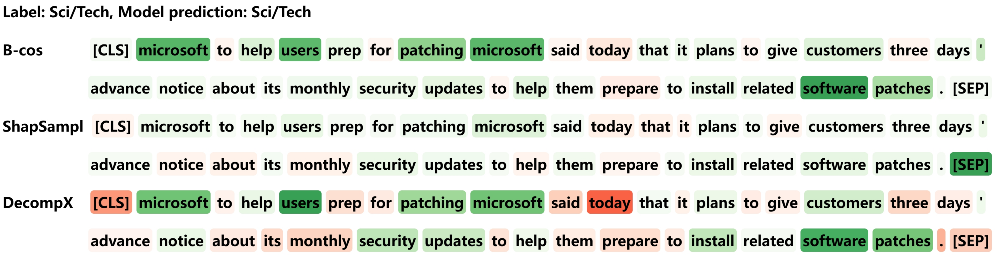

Label: Sci/Tech, Model prediction: Sci/Tech
B-cos
[CLS] microsoft to help users prep for patching microsoft said today that it plans to give customers three days '
advance notice about its monthly security updates to help them prepare to install related software patches . [SEP]
ShapSampl
[CLS] microsoft to help users prep for patching microsoft said today that it plans to give customers three days '
advance notice about its monthly security updates to help them prepare to install related software patches . [SEP]
DecompX
[CLS] microsoft to help users prep for patching microsoft said today that it plans to give customers three days '
advance notice about its monthly security updates to help them prepare to install related software patches . [SEP]

Figure 4: Examples of B-cos explanations (B-cos BERT) as well as ShapSampl and DecompX explanations (conv. BERT) from AG News. Green (red) indicates the positive (negative) impact of tokens on the prediction. The B-cos explanation highlights only relevant tokens and is more interpretable to humans (cf. Appendix I for more examples).

In Table 3, we find that removing alignment pressure (using B=1) degrades both task performance and explanation faithfulness. Replacing cross-entropy with BCE loss has little effect on classification accuracy, but improves faithfulness in perturbation-based evaluations. Architectural adaptations, including removing bias terms and eliminating activation functions in prediction heads, are also critical for enhancing task performance and explainability. Besides, we observe numerical instability when generating explanations without these architectural adaptations, as the dynamic linear weights for sigmoid and tanh ($\text{sigmoid}(\mathbf{x}) \times \mathbf{x}^{-1}$ and $\tanh(\mathbf{x}) \times \mathbf{x}^{-1}$) become unstable when $\mathbf{x}$ is close to zero.

In addition to ablations of model design and training components, we also evaluate alternative explanation methods. Replacing the dynamic linear weights $\mathbf{W}(\mathbf{x})$ with gradients (equivalent to IxG) yields less faithful explanations on B-cos LMs. Besides, directly extracting B-cos-like explanations, $\mathbf{W}(\mathbf{x})\mathbf{x}$, from a conventional model results in worse faithfulness compared to those from B-cos LMs.

	Acc (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)
Full system	78.64	59.66	-4.89	77.57
w/o alignment pressure (B=1)	78.07 (0.57)	57.19 (2.44)	-2.57 (2.32)	70.18 (7.39)
w/o BCE training	79.00 (0.36)	49.22 (10.44)	-7.91 (3.02)	79.21 (1.64)
w/o architectural adaptations	77.65 (0.99)	52.23 (7.43)	-3.80 (1.09)	74.30 (3.27)
w/o dynamic linear weights (IxG)	78.64 (0.00)	44.93 (14.73)	-0.60 (4.29)	53.57 (24.00)
$\mathbf{W}(\mathbf{x})\mathbf{x}$ from conv. model	80.77 (2.13)	44.92 (14.74)	2.80 (7.69)	70.20 (7.37)

Table 3: Ablation study of key designs in the B-cos BERT model on HateXplain. Values in parentheses indicate the difference from the full model's performance. Green (red) indicates the results are better (worse) than the full system.

5 Comparison of B-cosification Setups

Transforming PLMs into task-specific B-cos LMs involves two key objectives: task fine-tuning and B-cos conversion. While our main experiments combine these two phases, they can also be performed separately. To assess their effects, we compare two alternative training setups:

- Task then B-cos: PLMs are first fine-tuned on a downstream task. B-cos adaptations are then applied, followed by further fine-tuning on the same task for B-cos conversion. This setup is equivalent to Arya et al. (2024) who apply B-cosification to models with downstream task capabilities.
- B-cos then task: B-cos adaptations are applied to PLMs first, followed by pre-training on unsupervised texts to enhance B-cosification. The pre-trained B-cos models are then fine-tuned on the downstream task.

We evaluate these setups against the B-cosification approach used in our main experiments (B-cos LM) and compare task performance, faithfulness, and training efficiency (cf. Appendix C for B-cos pre-training details).

Additionally, we report results for conventional fine-tuning (Conv. LM) and training a randomly initialized B-cos LM (B-cos from scratch). Experiments are conducted on IMDB with B=1.25 for B-cos models, with results averaged over three runs.

Table 4 shows that B-cos LM requires fewer training steps to reach optimal validation performance than conventional fine-tuning. Training B-cos LM from scratch results in worse accuracy and faithfulness, emphasizing the importance of good parameter initialization. Among the two setups that separate task fine-tuning and B-cos conversion, *Task then B-cos* achieves results similar to B-cos LM but requires more total training steps. *B-cos then task* initially performs worse under the same training budget. However, with additional pre-training epochs, it surpasses other B-cosification setups in both task performance and faithfulness. Overall, we find that combining task fine-tuning and B-cos conversion is the most efficient approach. However, with sufficient pre-training, *B-cos then task* can produce more performant and explainable models.

Setup	**Epochs**	**Acc (↑)**	**SeqPG (↑)**	**Steps (K)**
Conv. LM	5	94.06	-	6.67
B-cos LM	5	91.00	70.66	4.33
B-cos from scratch	5	88.25	60.92	4.33
Task then B-cos	1+4	91.17	70.01	1+5
	2+3	91.30	70.48	3+3.33
	3+2	91.38	70.83	4+3
	4+1	89.56	70.66	5+1
	5+5*	91.27	70.78	6.67+3.33
B-cos then task	1+4	90.64	67.07	1+5
	2+3	91.04	68.97	3+4
	3+2	90.50	68.48	4+3
	4+1	89.18	69.92	6+1
	5+5*	91.45	71.86	7+5.33
	10+5*	92.19	73.44	15+6.33
	20+5*	92.87	75.01	31+6

Table 4: Training epochs, accuracy, explanation faithfulness, and convergence steps for different B-cosification setups. For two-phase methods, we report epoch distribution and convergence steps per phase. * marks additional training epochs.

6 Impact of B-cosification and B Values

For a deeper understanding of how B-cosification and alignment pressure parameter B affect model performance and behavior, we compare conventional and B-cos BERT trained on HateXplain across different B values. We also provide an empirical analysis of the impact of B on input-weight alignment in Appendix J.

Model Performance Figure 5 shows the effects of varying B on the task performance and explanation faithfulness. Classification accuracy initially improves slightly as B increases from 1 to 1.25, benefiting from the extra non-linearity introduced by B>1. However, beyond this point, accuracy declines as higher alignment pressure reduces model flexibility. A similar trend is observed for Comp, which peaks around B=1.5 before decreasing. This differs from previous findings in vision models (Böhle et al., 2022), which we attribute to the high sparsity of explanations at larger B values. As alignment pressure increases, fewer tokens receive attribution scores that are not close to zero, leading to poor token importance calibration and lower Comp scores. The effects of B on other metrics are similar and can be found in Appendix K.

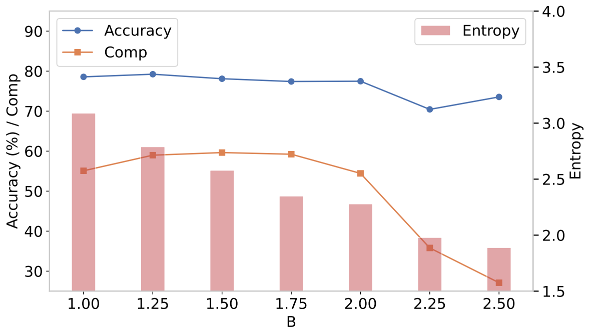


Figure 5: Varying B for B-cos BERT (HateXplain). Accuracy and Comp both peak around B=1.5, while explanation entropy negatively correlates with B.

Explanation Entropy Figure 5 also reveals a negative correlation between explanation entropy and B, indicating that higher alignment pressure leads to sparser explanations. This aligns with our expectations: a larger B amplifies the differences between dimensions in $|\cos(\mathbf{x}, \hat{\mathbf{W}})|^{\text{B-1}}$ of B-cos layers (Equation 2) and the dynamic linear weight assigns more distinct attributions to input features. As a result, explanations become more concentrated, where only a few tokens receive high attributions, while most remain close to zero (cf. Appendix L for an example).

Model Bias Since B-cos LMs with larger B values rely on fewer tokens for prediction, we investigate whether this may cause them to overfit and learn biases in the data. For this, we examine label bias and word-level spurious correlations using HateXplain, where approximately 60% of training and test examples have positive labels and societal biases are present. Figure 6 shows that a larger B value (B=2.5) reduces the model capacity, leading to a substantially higher positive rate in predictions and therefore lower class-balanced accuracy. Moreover, the B=2.5 model assigns higher attributions to non-semantic [CLS] and [SEP] tokens, indicating a reduced reliance on meaningful content. Notably, this label bias is not observed in the balanced AG News and IMDB datasets.

We also find that B-cosification, particularly with large B, amplifies reliance on spurious correlations. For example, the prediction positive rate for examples with the word "black" rises from 49.02% in the test set and 52.94% in the conventional model to 59.80%, 56.86%, and 73.53% in B-cos LMs with B=1, 1.5, and, 2.5, respectively (we provide an example in Appendix M). However, the faithfulness and interpretability of B-cos explanations facilitate the detection of spurious correlations and can effectively guide models toward reducing them (Rao et al., 2023). We leave the exploration of B-cos LMs for bias detection and mitigation to future work.

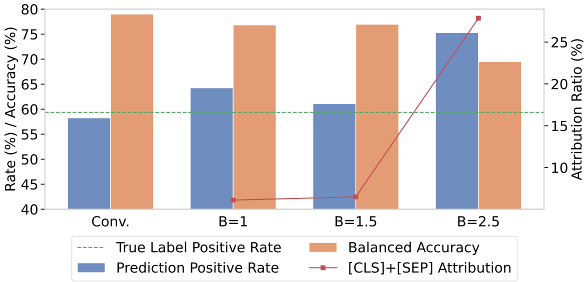


Figure 6: Comparison of conv. BERT and B-cos BERT with different B values. The attributions to [CLS] and [SEP] tokens (🟥) indicate that B-cos LMs with large B overfit to the non-semantic label distribution.

7 B-cosifying Decoder-Only Models for Generation Tasks

LLMs are increasingly used as general-purpose assistants, with most based on decoder-only architectures (Zhao et al., 2023; Minaee et al., 2024). While our primary focus is on classification tasks using encoder-only models, we also extend B-cosification to decoder-only models for generation tasks to demonstrate the broader applicability of B-cos LMs. Specifically, we apply B-cosification to two decoder-only models, GPT-2 small (Radford et al., 2019, referred to as GPT-2 afterwards) and Llama-3.2-1B (Dubey et al., 2024, referred to as Llama-3.2 afterwards), and evaluate their language modeling performance and explanation quality on two generation tasks. For more details on the datasets, experimental setup and baseline models, see Appendix C.

B-cosification Setup Given the complexity of modeling natural language, we use a small B value of 1.1. We do not B-cosify the language head, as its parameters are tied with the embedding layer. We use the standard cross-entropy loss instead of BCE, since the unnormalized language head weights could otherwise grow arbitrarily large to minimize the loss. To convert GPT-2 and Llama-3.2 to B-cos LMs, we apply B-cos adaptations and further train them on 500,000 and 4,000,000 sentences from OpenWebText[3], respectively.

Datasets For explanation evaluation, we use the BLiMP dataset (Warstadt et al., 2020) to assess explanations for linguistic phenomena, and the Indirect Object Identification (IOI) dataset (Brian Muhia, 2022) to test models' reasoning about object identification. Following Ferrando et al. (2023), we use nine subsets of BLiMP. Each example in both datasets consists of a sentence prefix followed by a target and a foil next-word prediction, differing in whether they align with the phenomenon or ability of interest. Ground truth evidence is provided to support either grammatical correctness or correct object identification. Examples of these datasets can be found in Table 7 in Appendix C.

Model	**Probability Gap (↑)**		**PPL (↓)**
	BLiMP	IOI	
GPT-2	0.0055	0.3351	3.10
B-cos GPT-2	**0.0059**	0.3265	**3.04**
Llama-3.2	0.0058	0.4652	2.51
B-cos Llama-3.2	**0.0065**	**0.5021**	2.64

Table 5: Language ability results for vanilla and B-cos decoder-only models. Scores where B-cos LM outperforms their vanilla counterparts are in **bold**. B-cos LMs show language modeling ability comparable to vanilla models. Results for each subset can be found in Table 13 in Appendix N.

[3] https://huggingface.co/datasets/Skylion007/openwebtext

Metrics and Baselines We evaluate explanation quality using Mean Reciprocal Rank (MRR), where higher scores indicate stronger alignment with the ground truth evidence. To assess language modeling abilities of models, we report two metrics: (1) the probability gap between target and foil predictions, and (2) perplexity (PPL) on a held-out corpus. Following Yin & Neubig (2022), we generate contrastive explanations that explain why the model predicts target tokens instead of foil tokens, and compare B-cos explanations against several baseline methods: L1 gradient norm (Grad Norm), IxG, Occlusion, and two propagation-based methods Logit and ALTI Logit from Ferrando et al. (2023).

Method	**GPT-2 MRR (↑)**		**Llama-3.2 MRR (↑)**	
	BLiMP	IOI	BLiMP	IOI
Random	0.5130	0.2360	0.5132	0.2328
Grad Norm	0.5465	0.8599	0.5504	0.3637
IxG	0.4750	0.1112	0.5303	0.1034
Occlusion	0.6365	0.8517	0.6201	0.4767
Logit	0.7307	**1.0**	-	-
ALTI Logit	0.7391	**1.0**	-	-
B-cos	**0.7561**	**1.0**	**0.6969**	**0.9913**

Table 6: Alignment results (MRR) on BLiMP and IOI. Logit and ALTI Logit results are replicated from the original paper (Ferrando et al., 2023). Best scores are marked in **bold**. B-cos explanations achieve the best alignment with ground truth evidence. Results for each subset can be found in Table 14 and Table 15 in Appendix N.

Results Table 5 shows that B-cos GPT-2 and B-cos Llama-3.2 models achieve strong language modeling performance comparable to their vanilla counterparts. Besides, Table 6 demonstrates that B-cos explanations exhibit better alignment with ground truth across tasks and models, indicating improved explainability of B-cos decoder-only LMs. Although the current B-cosification pipeline requires additional training, future work could explore more efficient approaches that reduce training overhead or integrate B-cosification into the pre-training phase. Overall, we believe B-cos decoder-only models are well-suited for tasks where explainability is critical and represent a promising direction for building more transparent and reliable LLMs.

8 Conclusion

In this work, we introduce B-cos LM, a bias-free dynamic linear model that learns task-relevant patterns through increased input-weight alignment pressure. B-cos LMs generate more faithful and human interpretable explanations while maintaining strong task performance and fast convergence. Based on our in-depth analysis of B-cosification, we provide three recommendations for effectively transforming PLMs into B-cos LMs: (1) combine B-cos conversion and task fine-tuning for efficient B-cosification. If resources allow, additional B-cos pre-training can further improve task performance and explanation faithfulness; (2) carefully select the parameter B, as excessively large values can reduce model capacity and lead to overly sparse explanations; and (3) be mindful of biases in training data, especially at high B values, as B-cosification may amplify existing biases.

We also explore adapting decoder-only models into B-cos LMs for generation tasks and show that, with additional training, they match the language modeling performance of conventional models while providing better explanations. We hope these findings support future efforts in building explainable LLMs.

9 Limitations

This study has certain limitations that should be acknowledged. First, the automatic evaluation metrics we use may not fully capture the faithfulness of different explanation methods (Feng et al., 2018; Lapuschkin et al., 2019). However, since there is no universal consensus on the most reliable evaluation metrics, this remains an open challenge in explainability research.

Second, we find that B-cos explanations do not consistently capture token interactions within multi-token phrases. For example, a negation phrase like *not good* tends to receive an overall attribution score that aligns with its meaning (e.g., a negative score for positive sentiment), but the individual token scores within the phrase vary across contexts. In some cases, the word *good* may receive either positive or negative scores across different examples, even when the overall sentiment remains the same. Similar issues arise in other methods, suggesting a broader limitation of token-level rationales in capturing compositional semantics.

References

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf`.

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. Evaluating recurrent neural network explanations. In Tal Linzen, Grzegorz Chrupał a, Yonatan Belinkov, and Dieuwke Hupkes (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 113–126, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4813. URL `https://aclanthology.org/W19-4813/`.

Shreyash Arya, Sukrut Rao, Moritz Böhle, and Bernt Schiele. B-cosification: Transforming deep neural networks to be inherently linterpretable. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 62756–62786. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/72d50a87b218d84c175d16f4557f7e12-Paper-Conference.pdf`.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3256–3274, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.263. URL `https://aclanthology.org/2020.emnlp-main.263/`.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Diagnostics-guided explanation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10445–10453, 2022.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1409.0473`.

Oren Barkan, Yonatan Toib, Yehonatan Elisha, Jonathan Weill, and Noam Koenigstein. LLM explainability via attributive masking learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9522–9537, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.556. URL `https://aclanthology.org/2024.findings-emnlp.556/`.

Moritz Böhle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10029–10038, 2021.

Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10329–10338, June 2022.

Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele. B-cos alignment for inherently interpretable cnns and vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas

Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936. doi: http://dx.doi.org/10.4135/9781412961288.n455.

Lola Le Breton, Quentin Fournier, Mariam El Mezouar, and Sarath Chandar. Neobert: A next-generation BERT. *CoRR*, abs/2502.19587, 2025. doi: 10.48550/ARXIV.2502.19587. URL https://doi.org/10.48550/arXiv.2502.19587.

Brian Muhia. ioi (revision 223da8b), 2022. URL https://huggingface.co/datasets/fahamu/ioi.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Antoine Chaffin. Reason-moderncolbert, 2025. URL https://huggingface.co/lightonai/Reason-ModernColBERT.

George Chrysostomou and Nikolaos Aletras. Enjoy the salience: Towards better transformer-based faithful explanations with word salience. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8189–8200, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.645. URL https://aclanthology.org/2021.emnlp-main.645/.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. Contextcite: Attributing model generation to context. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95764–95807. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/adbea136219b64db96a9941e4249a857-Paper-Conference.pdf.

Björn Deiseroth, Mayukh Deb, Samuel Weinbach, Manuel Brack, Patrick Schramowski, and Kristian Kersting. Atman: Understanding transformer predictions through memory efficient attention manipulation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 63437–63460. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c83bc020a020cdeb966ed10804619664-Paper-Conference.pdf.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7555–7565, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.477. URL https://aclanthology.org/2023.findings-acl.477/.

James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan Arik, Yan Liu, and Tomas Pfister. TextGenSHAP: Scalable post-hoc explanations in text generation with long documents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13984–14011, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.832. URL https://aclanthology.org/2024.findings-acl.832/.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3719–3728, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1407. URL https://aclanthology.org/D18-1407/.

Zijian Feng, Hanzhang Zhou, Zixiao Zhu, Junlang Qian, and Kezhi Mao. Unveiling and manipulating prompt influence in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ap1ByuwQrX.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. Explaining how transformers use context to build predictions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pp. 5486–5513, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.301. URL https://aclanthology.org/2023.acl-long.301/.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL http://arxiv.org/abs/1606.08415.

Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving deep learning interpretability by saliency guided training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26726–26739. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e0cd3f16f9e883ca91c2a4c24f47b3d9-Paper.pdf.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Alon Jacovi and Yoav Goldberg. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310, 2021. doi: 10.1162/tacl_a_00367. URL https://aclanthology.org/2021.tacl-1.18/.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Han Jiang, Junwen Duan, Zhe Qu, and Jianxin Wang. MARE: Multi-aspect rationale extractor on unsupervised rationale extraction. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11734–11745, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.655. URL https://aclanthology.org/2024.emnlp-main.655/.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. Investigating the influence of noise and distractors on the interpretation of neural networks. *CoRR*, abs/1611.07270, 2016. URL http://arxiv.org/abs/1611.07270.

Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and patternattribution. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Hkn7CBaTW.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 267–280, 2019.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896, 2020. URL https://arxiv.org/abs/2009.07896.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An evaluation of the human-interpretability of explanation. *CoRR*, abs/1902.00006, 2019. URL http://arxiv.org/abs/1902.00006.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1011. URL https://aclanthology.org/D16-1011/.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9119–9130, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.733. URL https://aclanthology.org/2020.emnlp-main.733/.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220, 2016. URL http://arxiv.org/abs/1612.08220.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 1–10. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723, June 2024. doi: 10.1162/coli_a_00511. URL https://aclanthology.org/2024.cl-2.6/.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *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 https://aclanthology.org/P11-1015/.

Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 295–337, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.19. URL https://aclanthology.org/2024.findings-acl.19/.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14867–14875, 2021.

Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *CoRR*, abs/2402.06196, 2024. doi: 10.48550/ARXIV.2402.06196. URL https://doi.org/10.48550/arXiv.2402.06196.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 258–271, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.19. URL https://aclanthology.org/2022.naacl-main.19/.

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. DecompX: Explaining transformers decisions by propagating token decomposition. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2649–2664, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.149. URL https://aclanthology.org/2023.acl-long.149/.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. Training with adversaries to improve faithfulness of attention in neural machine translation. In Boaz Shmueli and Yin Jou Huang (eds.), *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pp. 93–100, Suzhou, China, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.aacl-srw.14. URL https://aclanthology.org/2020.aacl-srw.14/.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. Measuring and improving faithfulness of attention in neural machine translation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2791–2802, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.243. URL https://aclanthology.org/2021.eacl-main.243/.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In Johannes Fürnkranz and Thorsten Joachims (eds.), *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pp. 807–814. Omnipress, 2010. URL https://icml.cc/Conferences/2010/papers/432.pdf.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Nina Poerner, Hinrich Schütze, and Benjamin Roth. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 340–350, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1032. URL https://aclanthology.org/P18-1032/.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. Learning to deceive with attention-based explanations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4782–4793, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.432. URL https://aclanthology.org/2020.acl-main.432/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Naveen Janaki Raman, Mateo Espinosa Zarlenga, and Mateja Jamnik. Understanding inter-concept relationships in concept-based models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42009–42025. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/raman24a.html.

Sahana Ramnath, Brihi Joshi, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, and Xiang Ren. Tailoring self-rationalizers with multi-reward distillation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=t8eO0CiZJV.

Sukrut Rao, Moritz Böhle, Amin Parchami-Araghi, and Bernt Schiele. Studying how to efficiently and effectively guide models with explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1922–1933, October 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pp. 1135–1144, 2016.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Punyajoy Saha, Divyanshu Sheth, Kushal Kedia, Binny Mathew, and Animesh Mukherjee. Rationale-guided few-shot classification to detect abusive language. In Kobi Gal, Ann Nowé, Grzegorz J. Nalepa, Roy Fairstein, and Roxana Radulescu (eds.), *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pp. 2041–2048. IOS Press, 2023. doi: 10.3233/FAIA230497. URL https://doi.org/10.3233/FAIA230497.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMlR, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6034.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL http://arxiv.org/abs/1706.03825.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6806.

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/80537a945c7aaa788ccfcdf1b99b5d8f-Paper.pdf.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Martin Tutek and Jan Šnajder. Toward practical usage of the attention mechanism as a tool for interpretability. *IEEE access*, 10:47011–47030, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupał a, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller, and Vera Schmitt. Cross-refine: Improving natural language explanation generation by learning in tandem. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1150–1167, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.77/.

Shengjie Wang, Tianyi Zhou, and Jeff Bilmes. Bias also matters: Bias attribution for deep neural network explanation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6659–6667. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/wang19p.html.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *CoRR*, abs/2412.13663, 2024. doi: 10.48550/ARXIV.2412.13663. URL https://doi.org/10.48550/arXiv.2412.13663.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Yeo Wei Jie, Ranjan Satapathy, and Erik Cambria. Plausible extractive rationalization through semi-supervised entailment signal. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5182–5192, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.307. URL https://aclanthology.org/2024.findings-acl.307/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Kayo Yin and Graham Neubig. Interpreting language models with contrastive explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 184–198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.14. URL https://aclanthology.org/2022.emnlp-main.14/.

Xuemin Yu, Fahim Dalvi, Nadir Durrani, Marzia Nouri, and Hassan Sajjad. Latent concept-based explanation of NLP models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12435–12459, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.692. URL https://aclanthology.org/2024.emnlp-main.692/.

Linan Yue, Qi Liu, Yichao Du, Yanqing An, Li Wang, and Enhong Chen. Dare: Disentanglement-augmented rationale extraction. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26603–26617. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a9a67d9309a28372dde3de2a1c837390-Paper-Conference.pdf.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Hang Zhao, Qile P. Chen, Yijing Barry Zhang, and Gang Yang. Advancing single- and multi-task text classification through large language model fine-tuning. *CoRR*, abs/2412.08587, 2024. doi: 10.48550/ARXIV.2412.08587. URL https://doi.org/10.48550/arXiv.2412.08587.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *CoRR*, abs/2303.18223, 2023. doi: 10.48550/ARXIV.2303.18223. URL https://doi.org/10.48550/arXiv.2303.18223.

Yiming Zheng, Serena Booth, Julie Shah, and Yilun Zhou. The irrationality of neural rationale models. In Apurv Verma, Yada Pruksachatkun, Kai-Wei Chang, Aram Galstyan, Jwala Dhamala, and Yang Trista Cao (eds.), *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pp. 64–73, Seattle, U.S.A., July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.trustnlp-1.6. URL https://aclanthology.org/2022.trustnlp-1.6/.

A Terminology

To ensure clarity, we define key terms used in this work as follows:

- **Faithfulness** The extent to which an explanation accurately reflects the model's actual reasoning process (Jacovi & Goldberg, 2020). A faithful explanation should directly correspond to the internal mechanisms that led to the model's prediction.
- **Human Interpretability** The ease with which a person can understand the model's reasoning from the explanation (Lage et al., 2019). A highly interpretable explanation should be clear, concise, and focused on relevant information while avoiding unnecessary or distracting information. However, an explanation that is easy for humans to interpret may not necessarily reflect the model's actual reasoning process or align with human reasoning patterns.
- **Human Agreement** The degree to which a model's explanation aligns with the reasoning a human would use for the same prediction. A high-agreement explanation should follow intuitive, logical reasoning patterns similar to human decision-making.
- **Explainability** The extent to which a model's computations can be faithfully explained and its learned patterns are understandable to humans. A highly explainable model should yield explanations that are both faithful to its actual reasoning process and interpretable to humans.

B Notation

In this paper, we use lowercase letters for scalars (e.g., b), bold lowercase letters for vectors (e.g., $\mathbf{w}$, $\mathbf{x}$), and bold uppercase letters ($\mathbf{W}$) for matrices.A special case is the alignment pressure parameter, denoted by the non-bold uppercase letter B, to distinguish it from the bias term b in linear layers. We use bold uppercase letters $\mathbf{X}$ and $\mathbf{A}$ to denote a sequence of model inputs or hidden state activations. In § 3, we use $\mathbf{x}$ to denote the input when a function is applied to each element of the input sequence separately. In contrast, we use $\mathbf{X}$ or $\mathbf{A}$ when the function involves interactions between elements, such as in the attention mechanism.

C Implementation Details

Fine-tuning Setups For all PLMs used in the experiments, we use the uncased base version from huggingface (Wolf et al., 2020). For both conventional models and B-cos LMs, we train them for 5 epochs with 10% linear warm-up steps on the downstream task datasets. The learning rates are set to 2e-5 for IMDB and HateXplain, and 3e-5 for AG News. All models use a batch size of 16 and a maximum sequence length of 512. For validation, we randomly sample half of the test set from IMDB and AG News.

Baselines For IxG and ShapSampl, we use the Captum (Kokhlikyan et al., 2020) implementations.[4] We implement the Attention method ourselves, and LIME is sourced from the lit library[5]. For DecompX[6] and SIG[7], we use their official implementations with default configurations. The number of samples is set to 25 for ShapSampl and 3,000 for LIME, with [MASK] as the baseline token. For all explanation methods at the embedding level, model predictions are attributed to the combined sum of word, position, and token type embeddings (if applicable). In the main experiments, we compute token attribution scores by summing over all embedding dimensions, as this approach demonstrates better faithfulness results than using the L2 norm.

For Saloss models, we use the official codebase[8] with default hyperparameters to train BERT and RoBERTa on AG News, IMDB, and HateXplain. DistilBERT is not included, as it is not supported by the codebase.

In Section 7, we follow Ferrando et al. (2023) to generate contrastive explanations that highlight why the models predicts the target token instead of the foil token. For Occlusion explanations, we use the [PAD] token to perform occlusion, instead of a zero vector as done in Yin & Neubig (2022) and Ferrando et al. (2023). Using zero vectors distorts the input distribution and, in generative settings, can influence predictions differently depending on position. To avoid such positional effects, we instead occlude using the in-distribution embedding of the [PAD] token.

SeqPG Examples When constructing examples for SeqPG, we set the sequence length to 50 for AG News, 256 for IMDB, and 25 for HateXplain, aligning with their median lengths. Only examples longer than these thresholds are selected, and they are truncated to construct synthetic examples. Additionally, we only use examples that are correctly predicted with a minimum confidence of 75% after truncation. For a fair comparison, we evaluate Saloss models and B-cos LMs on the same sets of examples constructed based on the predictions of the corresponding conventional models.

Automatic Evaluation Setups For task performance evaluation, we use the complete test set for each task. For faithfulness evaluation, we conduct perturbation-based evaluations on 2000 test examples and SeqPG on 500 test examples for AG News and IMDB. For HateXplain, we use the full test set for perturbation-based evaluation (1,924 examples) and construct 269, 310, and 308 SeqPG examples from it using BERT, DistilBERT, and RoBERTa, respectively. In the perturbation-based evaluation, the [CLS] token is never perturbed because it is used directly to make predictions.

B-cos Pre-training For B-cos pre-training in § 5, we set B=1.25 and further pre-train the model on 25,000 sentences from the Wikipedia dataset[9] using masked language modeling loss with a learning rate of 1e-4 and a 15% masking ratio. We do not B-cosify the language head, as its parameters are tied with the

[4] https://captum.ai/api/

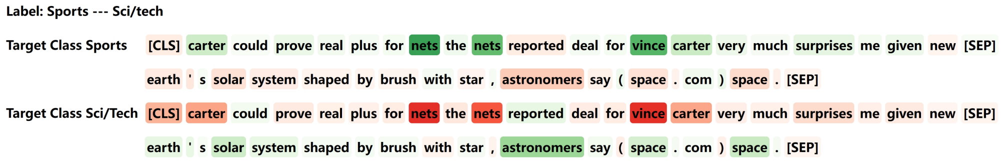
[5] https://github.com/PAIR-code/lit

[6] https://github.com/mohsenfayyaz/DecompX

[7] https://github.com/josephenguehard/time_interpret

[8] https://github.com/GChrysostomou/saloss

[9] https://huggingface.co/datasets/wikimedia/wikipedia


Label: Sports --- Sci/tech

Target Class Sports [CLS] carter could prove real plus for nets the nets reported deal for vince carter very much surprises me given new [SEP]
earth ' s solar system shaped by brush with star , astronomers say (space . com) space . [SEP]

Target Class Sci/Tech [CLS] carter could prove real plus for nets the nets reported deal for vince carter very much surprises me given new [SEP]
earth ' s solar system shaped by brush with star , astronomers say (space . com) space . [SEP]

Figure 7: An example of SeqPG from AG News (using B-cos BERT). Green (red) indicates the positive (negative) impact of tokens on the prediction. The example consists of two sequences with different labels (Sports and Sci/tech), separated by the [SEP] token after the first sequence. Explanations are generated for each label, and the proportion of correctly attributed positive tokens is averaged across both labels to compute the SeqPG score for this example.

embedding layer. Pre-training uses the standard cross-entropy loss rather than binary cross-entropy loss, since the unnormalized language head weights could otherwise grow arbitrarily large to minimize the loss.

Decoder-only Models B-cosification We use the GPT-2 small and Llama-3.2-1B models from huggingface. As with the encoder-based models, we do not B-cosify the language head and use cross-entropy loss for GPT-2 and Llama-3.2 training. B-cos adaptations are first applied and both models are then trained on 500,000 and 4,000,000 sentences, respectively, from the OpenWebText dataset for one epoch, using a learning rate of 5e-4. For GPT-2, we use a batch size of 16 and a maximum sequence length of 512; for Llama-3.2, we use a batch size of 128 and a sequence length of 1024. Perplexity is evaluated on a held-out OpenWebText subset of 10,000 sentences using a maximum sequence length of 512.

BLiMP Subsets We follow Ferrando et al. (2023) to use the following nine BLiMP subsets with corresponding IDs. aga: anaphor_gender_agreement; ana: anaphor_number_agreement; asp: animate_subject_passive; dna: determiner_noun_agreement_1; dnai: determiner_noun_agreement_irregular_1; dnaa: determiner_noun_agreement_with_adj_1; dnaai: determiner_noun_agreement_with_adj_irregular_1; npi: npi_present_1; darn: distractor_agreement_relational_noun. Examples of these datasets and the IOI dataset can be found in Table 7.

Dataset	ID	Example
Anaphor gender agreement	aga	Katherine can't help herself / himself.
Anaphor number agreement	ana	Susan revealed herself / themselves.
Animate subject passive	asp	Amanda was respected by some waitresses / pictures.
Determiner noun agreement 1	dna	Raymond is selling this sketch / sketches.
Determiner noun agreement irregular 1	dnai	Adam hadn't discussed these analyses / analysis.
Determiner noun agreement with adjectives 1	dnaa	Rebecca was criticizing those good documentaries / documentary.
Determiner noun agreement with adjectives irregular 1	dnaai	Some waiters broke this lost foot / feet.
NPI present 1	npi	Even Suzanne has really / ever joked around.
Distractor agreement relational noun	darn	A niece of most senators hasn't / haven't descended most slopes.
Indirect Object Identification	IOI	Friends Juana and Kristi found a mango at the bar. Kristi gave it to Juana / Kristi.

Table 7: Examples from the BLiMP and IOI datasets. Green (red) indicates target (foil) predictions. Ground truth evidence for the correct continuations is underlined.

Compute Infrastructure Unless stated otherwise, all experiments are conducted on a single NVIDIA H100 GPU. Training one epoch of B-cos BERT takes approximately 40 minutes on AG News, 10 minutes on IMDB, and 5 minutes on HateXplain.

D SeqPG Example

Figure 7 presents a SeqPG example from AG News using B-cos BERT. For better visualization, each segment is truncated to 20 tokens here instead of 50 used in the experiments. Unlike the hybrid document evaluation proposed by Poerner et al. (2018), our approach explicitly controls segment length and position to ensure a fair comparison. Additionally, we measure the proportion of correctly assigned positive attributions rather than relying solely on the highest attribution value.

E Task Performance of Other B-cos LMs

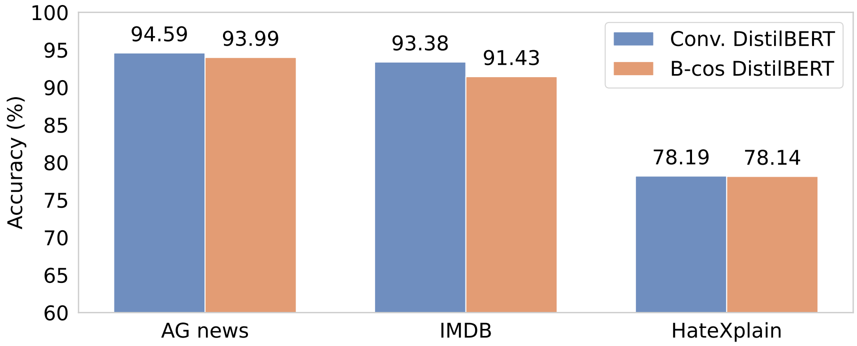


Figure 8: Mean accuracy of conventionally fine-tuned and B-cos DistilBERT models averaged over three runs. We use B=1.5, 1.25, and 1.5 for AG News, IMDB and HateXplain, respectively. B-cos models perform comparably to conventional models on most tasks.

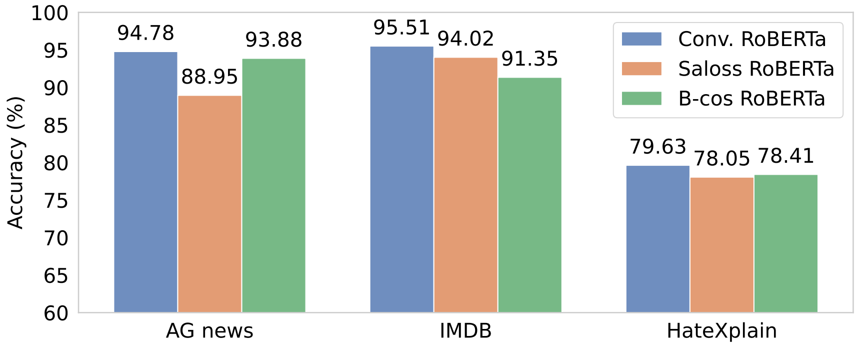


Figure 9: Mean accuracy of conventionally fine-tuned and B-cos RoBERTa models averaged over three runs. We use B=1.5, 1.25, and 1.5 for AG News, IMDB and HateXplain, respectively. B-cos models perform comparably to conventional models on most tasks.

Figures 8 and 9 illustrate the task performance of conventional and B-cos DistilBERT and RoBERTa across datasets. Consistent with findings from BERT models (cf. Figure 2), B-cos LMs exhibit strong performance comparable to conventionally fine-tuned models.

F Faithfulness Evaluation of Other B-cos LMs

Tables 8 and 9 present the faithfulness evaluation results for DistilBERT and RoBERTa. The findings are consistent with our main experiments (cf. Table 2), confirming that B-cos LMs produce more faithful explanations compared to post-hoc explanation methods.

Model	Method	AG News			IMDB			HateXplain		
		Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)
Conv. DistilBERT	Attention	26.36	5.37	50	31.62	10.46	50	30.56	14.67	50
	IxG	19.29	6.21	53.71	23.78	12.38	49.23	25.13	18.08	46.60
	SIG	30.78	1.63	67.87	47.16	5.48	60.66	41.11	4.23	58.55
	DecompX	-	-	-	-	-	-	-	-	-
	ShapSampl	52.56	-0.56	82.64	63.29	2.91	70.27	48.73	0.87	64.44
	LIME	52.59	-0.56	77.64	58.6	5.12	61.11	31.61	12.94	56.49
B-cos DistilBERT	B-cos	**61.93**	**-1.01**	**86.78**	**76.26**	**-1.28**	**72.68**	**57.2**	**-4.49**	**74.89**

Table 8: Faithfulness evaluation for conventionally fine-tuned DistilBERT and B-cos DistilBERT across three datasets. We use B=1.5, 1.25, and 1.5 for AG News, IMDB and HateXplain, respectively. The best results are in **bold**. We find that B-cos explanations are consistently more faithful than post-hoc explanations from both models.

Model	Method	AG News			IMDB			HateXplain		
		Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)
Conv. RoBERTa	Attention	22.17	3.80	50	25.26	5.84	50	32.94	7.52	50
	IxG	11.33	7.54	44.15	16.15	11.53	47.20	24.40	15.16	50.59
	SIG	19.64	1.63	66.43	38.14	2.13	59.04	44.21	-1.42	66.73
	DecompX	50.00	-0.84	**90.38**	49.24	0.65	72.80	46.94	-1.42	70.16
	ShapSampl	35.63	-0.68	78.31	43.32	1.83	65.85	44.83	-1.30	67.15
	LIME	19.28	2.85	66.73	21.07	8.32	50.81	27.97	11.38	58.59
Saloss RoBERTa	Attention	40.69	2.77	50	24.51	4.33	50	47.04	7.83	50
	IxG	6.19	27.30	52.46	10.98	12.30	47.92	22.78	25.78	49.49
	SIG	6.91	27.22	56.84	11.53	13.76	62.10	43.77	5.02	58.67
	DecompX	61.46	0.16	74.20	65.50	0.10	**74.41**	54.94	2.47	65.63
	ShapSampl	34.48	0.73	64.67	48.53	0.82	63.04	**55.80**	1.49	64.53
	LIME	15.93	8.03	55.17	18.04	6.47	50.94	29.62	15.78	56.00
B-cos RoBERTa	B-cos	**62.47**	**-1.18**	86.63	**73.87**	**-2.30**	74.05	51.33	**-5.18**	**74.01**

Table 9: Faithfulness evaluation for conventionally fine-tuned RoBERTa, Saloss RoBERTa and B-cos RoBERTa across three datasets. We use B=1.5, 1.25, and 1.5 for AG News, IMDB and HateXplain, respectively. The best results are in **bold**. We find that B-cos explanations are consistently more faithful than post-hoc explanations from both models.

G Comparison to Rationale-Based Models

We compare B-cos LMs to one rationale-based, explain-then-predict BERT model, RGFS-SA (Saha et al., 2023)[10] on HateXplain. This model leverages human rationales as additional supervision during training. As shown in Table 10, although the RGFS-SA model brings improvement over the conventional BERT model, it generates considerably less faithful rationales compared to B-cos explanations.

Model	**Method**	**Accuracy (↑)**	**Comp (↑)**	**Suff (↓)**	**SeqPG (↑)**
Conv. BERT	Attention	80.77	22.64	13.83	50
RGFS-SA BERT	Rationale	80.09	36.11	16.54	50
B-cos BERT	B-cos	78.64	59.66	-4.89	77.57

Table 10: Performance of conventional, RGFS-SA and B-cos (B=1.5) BERT models on HateXplain. SeqPG is consistently 50 for rationale-based models, as their explanations are class-agnostic, similar to attention. The rationale-based RGFS-SA model generates less faithful explanations than B-cos BERT.

[10] https://huggingface.co/Hate-speech-CNERG/Rationale_predictor

H Human Evaluation Details

In the human study, we select only examples shorter than 25 tokens for HateXplain and 40 tokens for AG News to improve visualization. Additionally, we replace [CLS] and [SEP] with ## to make the examples more understandable for lay users. Below, we provide the instructions along with a detailed description of the criteria and scoring used in our human evaluation. In our human study, 92% of AG News examples and 80% of HateXplain examples contain correct model predictions; in the remaining cases, explanations are supposed to support the wrong predictions.

> **WARNING: SOME CONTENT IN THIS QUESTIONNAIRE IS HIGHLY OFFENSIVE.**
>
> **Prerequisites:** Proficiency in English is required for this evaluation task. If you do not meet this criterion, please do not proceed.
>
> We invite you to review 100 examples where LMs perform classification tasks and provide explanations for their predictions.
>
> - The first 50 examples come from a hate speech detection task, where the model predicts whether a text is toxic or not toxic.
> - The last 50 examples come from a topic classification task, where the model categorizes a text into one of four topics: sports, world, business, or sci/tech.
>
> For each example:
>
> - The model's prediction is shown along with four explanations justifying the prediction.
> - The order of the explanations is randomized to prevent bias.
> - Words highlighted in green indicate words that had a positive influence on the prediction, while words in red indicate words that had a negative influence. The intensity of the color reflects the strength of the impact.
> - **Important:** The model's prediction may be incorrect. Your task is to evaluate the explanations based on how well they support the model's prediction, not the true labels.
>
> Evaluation Task:
>
> After reviewing each example, please rate the the **human interpretability** and **human agreement** of the four explanations on a scale of 1 to 5. Refer to the definitions and rating scales provided below when making your assessments.
>
> **Human Interpretability:** How easily a person can **understand the model's reasoning** based on the explanation. A highly interpretable explanation should be clear and easy to follow, focus on relevant words and avoid unnecessary or distracting details.
>
> 1. **Not Interpretable:** The explanation is unclear, noisy, or provides no meaningful insight.
> 2. **Slightly Interpretable:** Some clues are present, but the explanation is too sparse, irrelevant, or confusing.
> 3. **Moderately Interpretable:** The explanation contains useful information but is cluttered with noise or irrelevant details.
> 4. **Highly Interpretable:** The explanation is mostly clear, with minimal irrelevant highlights.
> 5. **Completely Interpretable:** The explanation is fully transparent, highlighting only the most relevant words, making the model's reasoning fully clear.
>
> **Human Agreement:** How closely the model's explanation **aligns with the reasoning a human would use** for the same prediction. A high-agreement explanation should follow logical, intuitive reasoning and align with typical human decision-making patterns.
>
> 1. **No Agreement:** The explanation contradicts human reasoning or lacks logic.
> 2. **Low Agreement:** The explanation bears some resemblance to human reasoning but includes major inconsistencies.

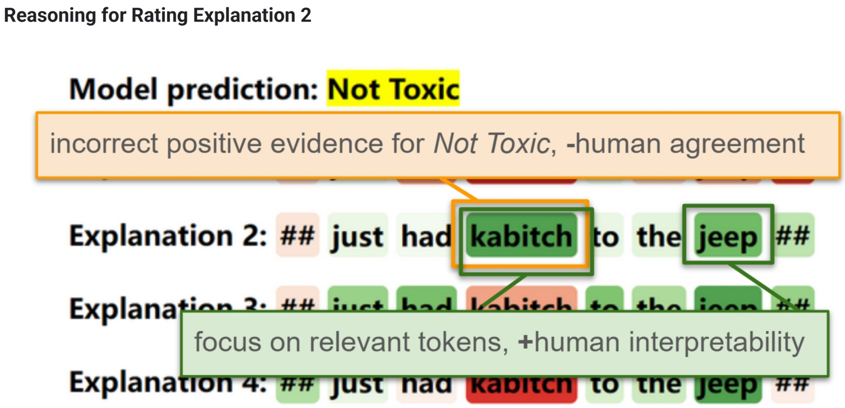
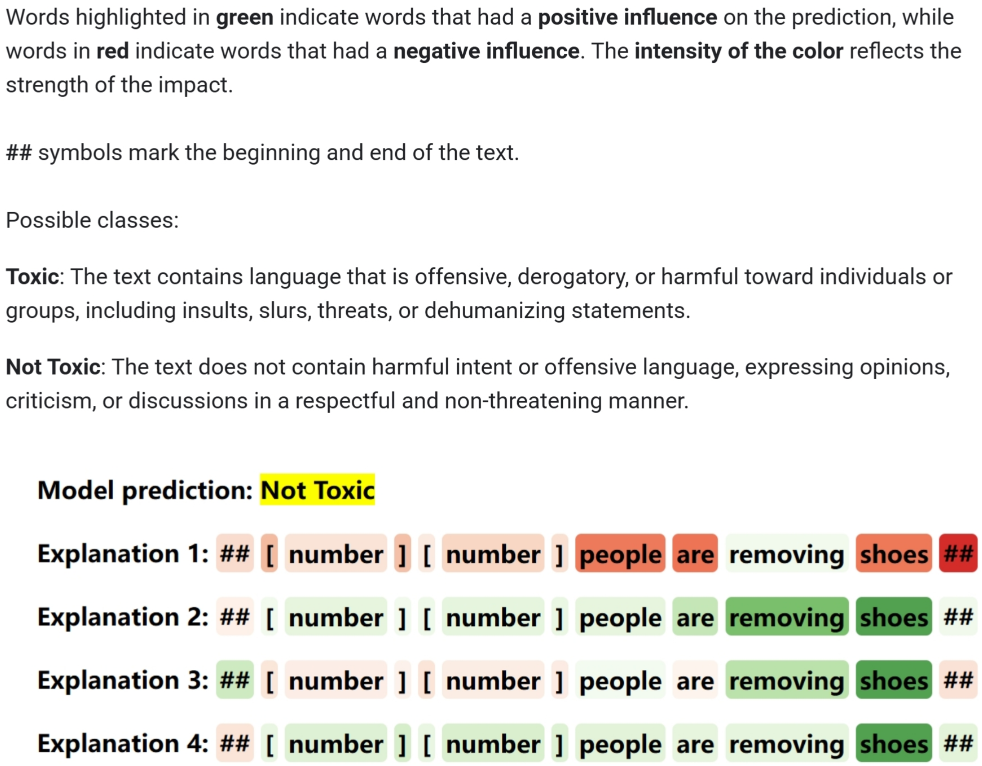

Reasoning for Rating Explanation 2
Model prediction: Not Toxic
incorrect positive evidence for *Not Toxic*, -human agreement
Explanation 2: ## just had kabitch to the jeep ##
Explanation 3: ## just had kabitch to the jeep ##
focus on relevant tokens, +human interpretability
Explanation 4: ## just had kabitch to the jeep ##

Figure 10: An example shown to participants that demonstrates how to rate explanations.

3. **Moderate Agreement:** The explanation partially aligns with human reasoning, yet contains notable differences.
4. **High Agreement:** The explanation largely aligns with human reasoning, showing only minor discrepancies.
5. **Complete Agreement:** The explanation fully matches human reasoning, following a logical and intuitive path that a human would naturally use.

We also provide participants with examples to illustrate the reasoning behind rating explanations. One such example is shown in Figure 10. Additionally, Figure 11 presents an example of a model prediction and its explanations as displayed to participants during the study.



Words highlighted in **green** indicate words that had a **positive influence** on the prediction, while words in **red** indicate words that had a **negative influence**. The **intensity of the color** reflects the strength of the impact.

symbols mark the beginning and end of the text.

Possible classes:

Toxic: The text contains language that is offensive, derogatory, or harmful toward individuals or groups, including insults, slurs, threats, or dehumanizing statements.

Not Toxic: The text does not contain harmful intent or offensive language, expressing opinions, criticism, or discussions in a respectful and non-threatening manner.

Model prediction: Not Toxic

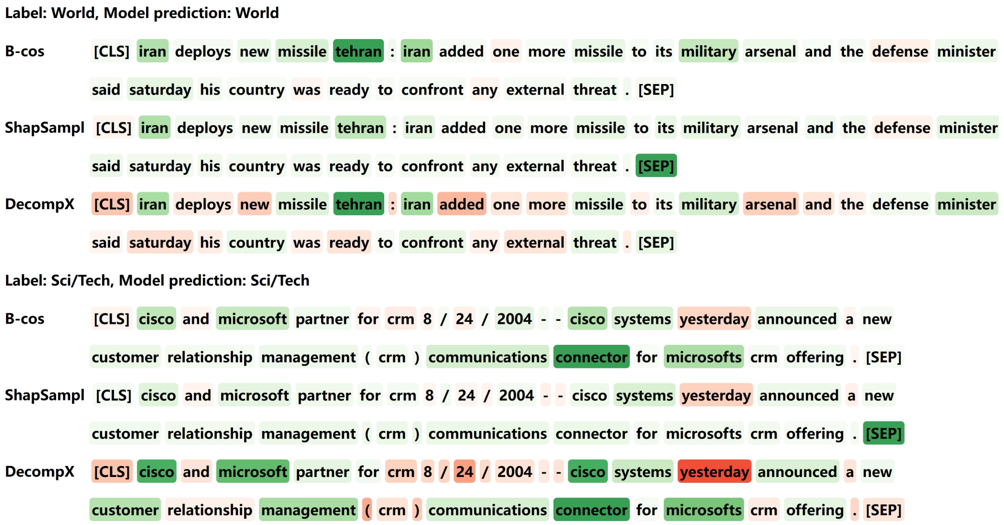
Explanation 1: ## [number] [number] people are removing shoes ##

Explanation 2: ## [number] [number] people are removing shoes ##

Explanation 3: ## [number] [number] people are removing shoes ##

Explanation 4: ## [number] [number] people are removing shoes ##

Figure 11: An examples of a model prediction and its explanations presented to participants.



Label: World, Model prediction: World

B-cos [CLS] iran deploys new missile tehran : iran added one more missile to its military arsenal and the defense minister said saturday his country was ready to confront any external threat . [SEP]

ShapSampl [CLS] iran deploys new missile tehran : iran added one more missile to its military arsenal and the defense minister said saturday his country was ready to confront any external threat . [SEP]

DecompX [CLS] iran deploys new missile tehran : iran added one more missile to its military arsenal and the defense minister said saturday his country was ready to confront any external threat . [SEP]

Label: Sci/Tech, Model prediction: Sci/Tech

B-cos [CLS] cisco and microsoft partner for crm 8 / 24 / 2004 - - cisco systems yesterday announced a new customer relationship management (crm) communications connector for microsofts crm offering . [SEP]

ShapSampl [CLS] cisco and microsoft partner for crm 8 / 24 / 2004 - - cisco systems yesterday announced a new customer relationship management (crm) communications connector for microsofts crm offering . [SEP]

DecompX [CLS] cisco and microsoft partner for crm 8 / 24 / 2004 - - cisco systems yesterday announced a new customer relationship management (crm) communications connector for microsofts crm offering . [SEP]

Figure 12: More examples of B-cos explanations (B-cos BERT) as well as ShapSampl and DecompX explanations (BERT) from the AG News dataset. Green (red) indicates the positive (negative) impact of tokens on the prediction. As can be seen, the B-cos explanation highlights only relevant tokens and is more interpretable to humans.

I More Examples of B-cos Explanations

We provide two more examples of B-cos and other (post-hoc) explanations from AG News in Figure 12. Consistent with our findings in § 4.4, B-cos LMs provide more human interpretable explanations.

I.1 Explanation Efficiency

Beyond improved faithfulness and human interpretability, B-cos explanations are also efficient to extract. Comparing their computational costs with strong post-hoc methods shows that B-cos explanations are the most efficient in both time and memory usage (Table 11). Post-hoc and B-cos explanations are generated from the conventionally fine-tuned and B-cos BERT models on IMDB, respectively.

J Impact of B on Input-weight Alignment

To analyze how B-cosification and alignment pressure influence the behavior of B-cos LMs, we compute the alignment (cosine similarity) between each input and its corresponding weight in B-cos modules across all layers. This analysis is performed on 100 examples from the HateXplain dataset. In Figure 13, we plot different percentiles of input-weight alignment for conventional and B-cos BERT models with varying B values. For better visualization, we display only the 10th to 90th percentiles.

Method	Time (s)	Memory (GB)
ShapSampl	37.22	21.95
LIME	6.82	21.96
SIG	67.46	29.09
DecompX	0.76	48.38
B-cos	**0.08**	**2.82**

Table 11: Computational costs per example of generating explanations for 100 instances using an NVIDIA H100 GPU (batch size 1). B-cos explanations (**bold**) are at least 9x faster and require at most $\frac{1}{8}$ of VRAM.

Overall, larger B values generally lead to stronger input-weight alignment compared to smaller B and conventional models, as evidenced by the curves for B=1.5 and B=2.5 lying above those for the conventional model and B=1. However, the alignment pattern becomes more complex when comparing B=1.5 and B=2.5. Specifically, at B=2.5, the most aligned input-weight pairs exhibit higher alignment than in other models, but some pairs show very low alignment. This result may arise because certain weights are highly optimized for specific input patterns, leading to poor alignment with others, particularly in later layers where input features become more anisotropic (Ethayarajh, 2019; Li et al., 2020). As a result, some outputs from the B-cos layers are highly

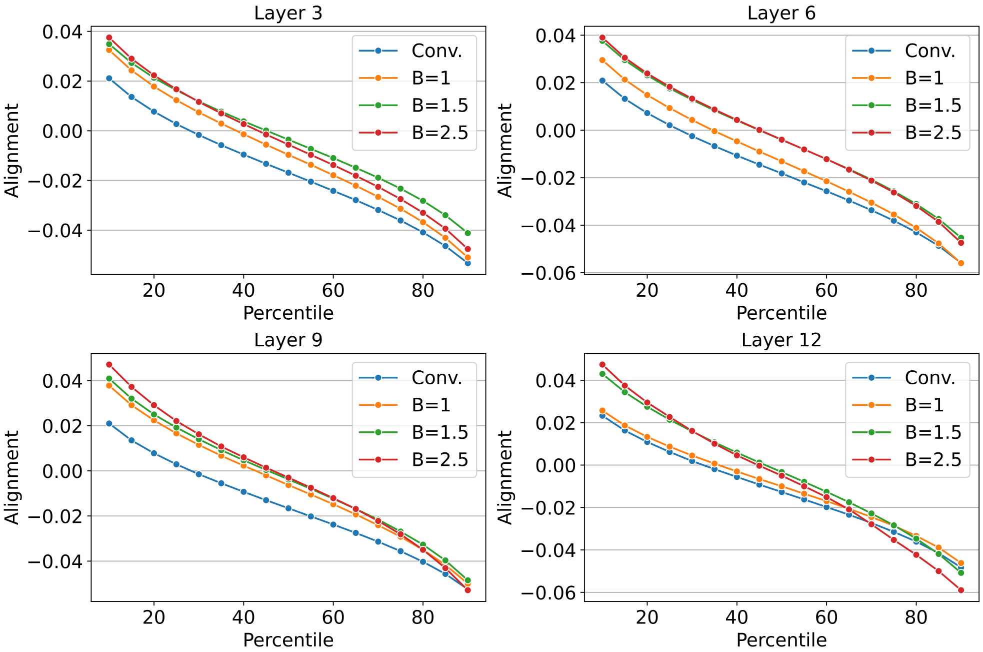


Figure 13: Percentiles of input-weight alignment in B-cos modules across selected layers of conventional and B-cos BERT models with different B values (HateXplain).

negative. When these outputs are fed into GELU activation functions, their dynamic weights approach zero, making the explanations more sparse.

K Effects of B on Other Metrics

Table 12 presents the complete results on how B values affect task performance, explanation faithfulness and explanation entropy, as shown in Figure 5. Similar to Comp, SeqPG scores also decline with higher alignment pressure. This could also be attributed to the high sparsity of explanations. As B increases, fewer tokens receive attribution scores that are not close to zero, and in some SeqPG examples, B-cos LMs may attribute predictions to a single segment. This can lead to numerical instability when computing the positive attribution ratio.

B	1.00	1.25	1.50	1.75	2.00	2.25	2.50
Acc (↑)	78.57	**79.23**	78.10	77.41	77.48	70.44	73.55
Comp (↑)	55.09	58.99	**59.64**	59.23	54.44	35.80	27.11
Suff (↓)	-4.25	-5.71	-5.47	-5.84	-6.69	**-7.23**	-5.47
SeqPG (↑)	69.75	77.26	**77.79**	77.67	76.79	76.68	77.25
Entropy	3.09	2.79	2.58	2.35	2.28	1.98	1.89

Table 12: Task performance, explanation faithfulness, and explanation entropy of B-cos BERT models on HateXplain with different B values. Results are averaged over three runs. Similar to Figure 5, task performance and explanation faithfulness peak around B=1.5, while explanation entropy negatively correlates with B.

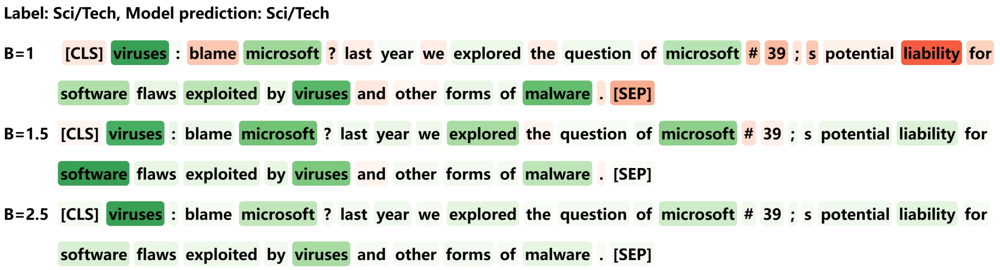


Figure 14: B-cos explanations (B-cos BERT) on AG News with different B values. Green (red) indicates the positive (negative) impact of tokens on the prediction. As B increases, B-cos LMs produce sparser explanations, with fewer tokens receiving significant attribution scores.

L B-cos Explanations with Different B Values

Figure 14 illustrates that with increased alignment pressure, B-cos LMs learn fewer but more task-relevant features. Consequently, they produce sparser explanations, with fewer tokens receiving significant attribution. This finding aligns with the statistics presented in § 6.

M Example of Model Bias

In the example shown in Figure 15, models become increasingly confident in the incorrect prediction as B increases, with attributions primarily assigned to the word "blacks". Moreover, simply replacing "blacks" with "whites" results in a sharp drop in confidence, which demonstrates a growing reliance on spurious correlations with increased alignment pressure. The observation further confirms our findings in §6.

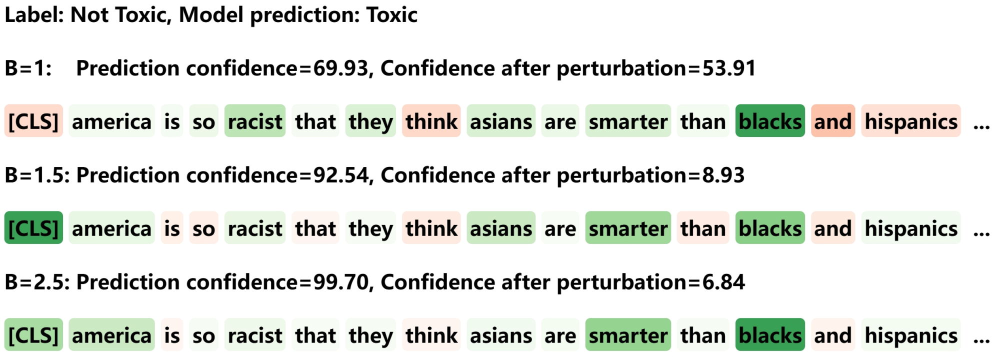


Figure 15: Example of how larger B values lead B-cos LMs to learn word-level spurious correlations. Green (red) indicates the positive (negative) impact of tokens on the prediction. Higher alignment pressure increases the reliance of B-cos LMs on spurious correlations in the data. In this example, perturbation involves changing "blacks" to "whites".

N Decoder-Only Model Results

Table 13 presents the average probability gaps between target and foil predictions on every dataset from different vanilla and B-cos models. Table 14 and Table 15 contain MRR results on every dataset in BLiMP and IOI for GPT-2 and Llama-3.2 models, respectively.

Dataset	Probability Gap (↑)			
	Vanilla GPT-2	B-cos GPT-2	Vanilla Llama-3.2	B-cos Llama-3.2
aga	0.0120	0.0170	0.0171	0.0196
ana	0.0152	0.0189	0.0140	0.0170
asp	0.0007	0.0008	0.0016	0.0009
dna	0.0011	0.0011	0.0011	0.0017
dnai	0.0021	0.0006	0.0011	0.0013
dnaa	0.0014	0.0012	0.0012	0.0017
dnaai	0.0091	0.0058	0.0077	0.0072
npi	0.0015	0.0002	0.0002	0.0001
darn	0.0067	0.0078	0.0080	0.0085
IOI	0.3351	0.3265	0.4652	0.5021

Table 13: Probability gaps between target and foil next token predictions from vanilla models and B-cos LMs on every dataset.

Dataset	Random	Grad Norm	IxG	Occlusion	Logit	ALTI Logit	B-cos
aga	0.6875	0.7927	0.7910	0.7513	0.827	0.964	0.8764
ana	0.7056	0.6753	0.7387	0.5957	0.817	0.976	0.7532
asp	0.3818	0.7512	0.4086	0.4374	0.386	0.499	0.4939
dna	0.4608	0.3629	0.3869	0.9030	0.737	0.646	0.9308
dnai	0.4626	0.4077	0.4317	0.8395	0.711	0.637	0.8596
dnaa	0.4103	0.2632	0.3214	0.6557	0.951	0.807	0.7798
dnaai	0.4074	0.2632	0.3392	0.6167	0.9	0.757	0.7601
npi	0.6121	0.7854	0.4948	0.4775	0.445	0.417	0.4573
darn	0.4888	0.6170	0.3627	0.4247	0.802	0.949	0.8936
IOI	0.2360	0.8599	0.1112	0.8517	1.0	1.0	1.0

Table 14: MRR Alignment of different explanation methods on GPT-2 small predictions on every dataset. B-cos explanations are extracted from the B-cos GPT-2 model. Logit and ALTI Logit results are duplicated from Ferrando et al. (2023).

Dataset	Random	Grad Norm	IxG	Occlusion	B-cos
aga	0.6868	0.6030	0.5928	0.7811	0.8485
ana	0.7037	0.5955	0.6432	0.6072	0.8535
asp	0.3842	0.7694	0.4537	0.3670	0.6108
dna	0.4615	0.4598	0.4898	0.8352	0.6743
dnai	0.4630	0.4542	0.5043	0.7671	0.6652
dnaa	0.4112	0.4299	0.4750	0.6115	0.6308
dnaai	0.4075	0.4221	0.4498	0.5758	0.5563
npi	0.6123	0.6367	0.7062	0.5154	0.6264
darn	0.4884	0.5828	0.45787	0.5210	0.8065
IOI	0.2328	0.3637	0.1034	0.4767	0.9913

Table 15: MRR Alignment of different explanation methods on Llama-3.2 predictions on every dataset. B-cos explanations are extracted from the B-cos Llama-3.2 model. As Llama models are not supported in Ferrando et al. (2023), we do not include their results.