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

Anonymous ACL submission

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 models. Meanwhile, B-cos networks have been introduced to improve model explainability through architectural and computational adaptations, but 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 networks empowered for 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 B-cos 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 provide practical guidelines for effectively building B-cos LMs based on our findings. Our code is available at https://anonymous.4open.science/r/bcos_lm.

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 explaining model predictions, particularly in relation to the input. These input-based explanations, often referred to as local explanations or rationales, aim to reveal how specific inputs influence a model's predictions (Arras et al., 2019; Atanasova et al., 2020; Lyu et al., 2024).

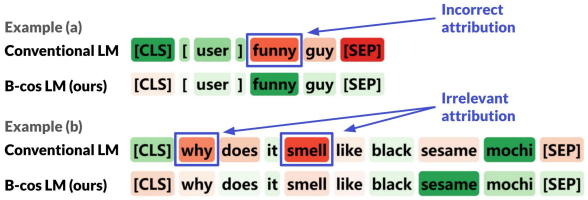

Figure 1: **Visualization of $\mathbf{W}(\mathbf{x})\mathbf{x}$ in a conventionally fine-tuned model (Conventional LM) and a B-cos LM.** Green (red) indicates the positive (negative) impact of tokens on the prediction. In both examples, both models correctly predict *not toxic*. In the Conventional LM,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

Most 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 and less interpretable explanations (Smilkov et al., 2017; Kindermans et al., 2019; Slack et al., 2020; Pruthi et al., 2020).[1] Prior research has attributed these shortcomings to the lack of explainability in contemporary neural models (Kindermans et al., 2018; Alvarez Melis and Jaakkola, 2018; Rudin, 2019). Figure 1 provides examples illustrating this issue.

To overcome these limitations, we introduce **B-cos LM**, a dynamic linear model that learns task-relevant patterns through increased input-weight alignment pressure. Building upon B-cos networks from computer vision (Böhle et al., 2022; Arya et al., 2024), we improve explainability of B-cos LMs through mathematically grounded architec-

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

tural and computational adaptations. Furthermore, we tailor B-cos LMs for NLP by incorporating specialized architectural modifications and training pipelines. 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. Based on our findings, we provide practical guidelines for building effective B-cos LMs.

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), propagation-based (Bach et al., 2015; Shrikumar et al., 2017; Springenberg et al., 2015), and perturbation-based methods (Li et al., 2016; Ribeiro et al., 2016; Lundberg and Lee, 2017). Besides, the attention mechanism (Bahdanau et al., 2015) is often viewed as an explanation, particularly in transformer-based models (Vaswani et al., 2017).

Although post-hoc methods can be applied to generate explanations for existing models, 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 and Wallace, 2019; Slack et al., 2020; Pruthi et al., 2020). Furthermore, they may generate noisy explanations that focus on irrelevant information, making them difficult for humans to interpret (Smilkov et al., 2017; Ismail et al., 2021).

From Post-hoc Explanations to Explainable Models The limitations of post-hoc explanation methods may be attributed to the inherent lack of explainability in contemporary neural models, which are typically optimized solely for task performance (Kindermans et al., 2018; Rudin, 2019; Atanasova et al., 2022). For instance, studies have shown that existing models struggle to provide faithful explanations (Alvarez Melis and Jaakkola, 2018) or tend to learn noisy patterns, resulting in less interpretable explanations (Ismail et al., 2021).

In response, various efforts have been made to enhance model explainability. Some work has introduced constraints that improve specific explanation properties, such as faithfulness (Tutek and Šnajder, 2022; Moradi et al., 2020, 2021), consistency (Atanasova et al., 2022), locality (Alvarez Melis and 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 and 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.

Property	Conventional Fine-tuning	B-cosification (Arya et al., 2024)	B-cos LM (ours)
Bias terms	yes	no	no
B (alignment pressure)	1	2	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 in computer vision and B-cosification in NLP (B-cos LM). Conventional fine-tuning and B-cosification follow the configuration of BERT for sequence classification and CLIP (Radford et al., 2021), respectively (cf. § 3 for details).

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 then describe how we transform PLMs to task-specific B-cos LMs (§ 3.2). Finally, we demonstrate how to generate explanations from B-cos LMs (§ 3.3). 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 and Hinton, 2010; Alvarez Melis and Jaakkola, 2018; Srinivas and Fleuret, 2019). For each input $\mathbf{x}$, the network 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 and Jaakkola, 2018). Earlier work refers to such models as dynamic linear models (Böhle et al., 2021; Böhle et al., 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. However, two challenges hinder the direct use of this interpretation. First, $\mathbf{W}(\mathbf{x})$ alone provides an incomplete and unfaithful explanation 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, $\mathbf{W}(\mathbf{x})$ is often difficult for humans to interpret, as it 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:

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

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 $\mathbf{w}(\mathbf{x}) = |\cos(\mathbf{x}, \hat{\mathbf{w}})|^{\mathbf{B-1}} \times \hat{\mathbf{w}}$. 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 3.3. Moreover, with additional alignment pressure (B>1), the weight $\mathbf{w}$ is forced to align closely with task-relevant patterns to achieve a high cosine similarity and strong activation within the B-cos module. As a result, only the most relevant features are highlighted in explanations, making them more interpretable to humans.

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 NLP models is non-trivial and inefficient due to the significant differences in model architectures and training pipelines.

3.2 B-cosification in NLP

In this section, we present our B-cosification approach for NLP. We summarize the differences be-

[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.

tween B-cosification in NLP, its counterpart in vision, and conventional fine-tuning in Table 1. We provide an extensive ablation study in Appendix C.

3.2.1 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 NLP models, we remove all activation functions in the prediction heads, as they reduce the locality of explanations and introduce numerical instability during their generation. We expect the added non-linearity from B>1 to compensates for this removal.

Introducing B-cos Computation To promote input-weight alignment and improve human interpretability of explanations, we follow Arya et al. (2024) and 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 $\mathbf{W}$ of the original model.

3.2.2 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 NLP-typical "pre-train then fine-tune" paradigm, 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 in NLP, as the objective involves both B-cos conversion and task fine-tuning. While there are multiple ways to conjoin these two steps (cf. § 7), 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 conventional cross-entropy loss, as it explicitly maximizes the absolute target logits and strengthens alignment pressure. We provide an extensive comparison of B-cosification setups in § 7.

3.3 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. Appendix D), the entire model computation can be expressed as a sequence of dynamic linear transformations:

$$\hat{\mathbf{W}}_L(\mathbf{A}_L)\hat{\mathbf{W}}_{L-1}(\mathbf{A}_{L-1})...\hat{\mathbf{W}}_1(\mathbf{A}_1 = \mathbf{X})\mathbf{X} \quad (3)$$

which can be completely summarized as a single dynamic linear function $\Pi_{j=1}^{L}\hat{\mathbf{W}}_j(\mathbf{A}_j)$.[3] Considering the textual inputs specific to NLP, 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 predicted 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 against conventional models and baseline explanation methods across various tasks, PLMs, and metrics.

Datasets and Models Our experiments include three sequence classification 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). We use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019) as the basis for conventional fine-tuning and for obtaining B-cos LMs (B=1.5) with the same training hyperparameters (cf. Appendix E for details on fine-tuning, B-cosification, and data splits).

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

[3] 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 $(\mathbf{W}(\mathbf{x}) + \mathbf{I}_n)\mathbf{x}$.

Model	Method	AG News			IMDB			HateXplain		
		Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)
(a) 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
(b) B-cos BERT	Attention	33.51	2.71	50	32.91	5.31	50	39.3	3.93	50
	IxG	26.89	1.31	48.24	61.74	-2.39	52.95	44.93	-0.60	53.57
	SIG	14.39	4.65	19.64	29.06	3.11	56.82	35.04	1.96	60.23
	DecompX	-	-	-	-	-	-	-	-	-
	ShapSampl	15.90	3.91	52.71	35.95	0.66	52.14	39.6	0.66	65.02
	LIME	57.99	0.07	79.30	70.05	-1.55	60.03	40.84	3.84	59.14
(c) B-cos BERT	B-cos	**64.22**	**-1.26**	**87.92**	**75.33**	**-2.95**	70.27	**59.66**	**-4.89**	**77.57**

Table 2: Faithfulness evaluation for conventionally fine-tuned BERT and B-cos BERT across three datasets. The best results are in **bold**. We find that B-cos explanations are consistently more faithful than post-hoc explanations from both models.

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 F.

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 and Kononenko, 2010), and LIME (Ribeiro et al., 2016). For a fair comparison against embedding-level methods, we aggregate attributions by summing across all embedding dimensions (cf. Appendix E).

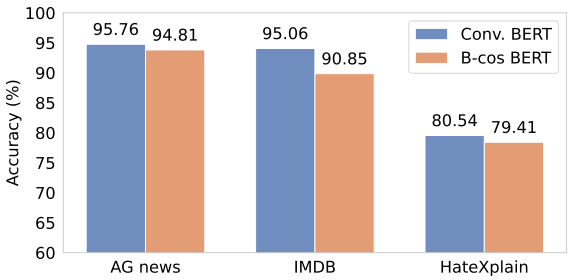

Figure 2: Mean accuracy of conventionally fine-tuned and B-cos BERT models averaged over three runs. B-cos models perform comparably to conventional models on most tasks.

Task Performance Figure 2 shows the accuracy of conventionally fine-tuned and B-cos BERT across three datasets (we provide results for DistilBERT and RoBERTa in Appendix G). We find that

all B-cos LMs performs on par with conventional models on AG News and HateXplain, with only a minor drop (∼1%) in accuracy. Only for IMDB, we find a slightly larger drop of 4.21%, though the performance remains strong overall.

Faithfulness Results Table 2 shows the faithfulness scores for post-hoc explanation methods on (a) conventionally fine-tuned BERT models and (b) B-cos BERT models, as well as (c) B-cos explanations extracted from B-cos BERT. The results show that B-cos explanations are consistently and substantially more faithful than post-hoc methods across all datasets. This improvement holds both across different models and within the same model. B-cos explanations outperform the strongest post-hoc methods on conventional models by an average of 14.63 points in Comp score and achieve negative Suff scores, indicating that the identified important tokens alone enable even more confident predictions. Additionally, B-cos explanations show a considerable improvement in SeqPG. Similar trends are observed for DistilBERT and RoBERTa (Appendix H), further strengthening our findings.

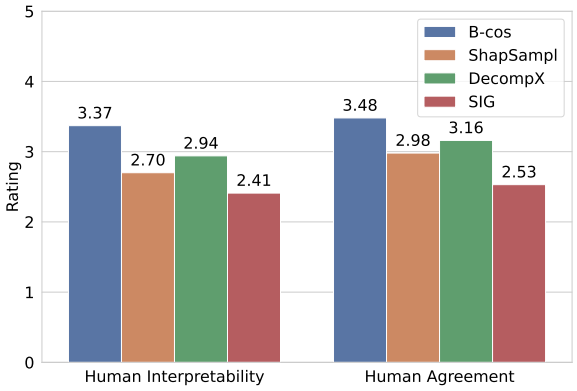

Figure 3: Human evaluation reveals that B-cos explanations have better human interpretability and human agreement than baseline methods and the improvements are statistically significant.

5 Human Evaluation

We conduct a human study to evaluate the human interpretability and agreement of B-cos explanations, comparing them against three strong post-hoc explanation methods on the conventional BERT model. For the study, we randomly select 50 instances from AG News and HateXplain where the B-cos and conventional model predict the same label. We then ask five annotators to rate the respective explanations in terms of human interpretability (how well they understand it) and human agreement (how much they agree with the it) on a scale of 1-5. We provide further details of the human evaluation in Appendix I.

Human Evaluation Results Figure 3 shows that B-cos explanations have a better human interpretability and exhibit greater alignment with human reasoning than post-hoc methods. Conducting paired t-tests with a Bonferroni-corrected $\alpha = \frac{0.05}{6} = 0.008\overline{3}$ (Bonferroni, 1936) shows that the improvements of B-cos explanations are statistically significant ($p < \alpha$) for both metrics.

6 Qualitative Analysis

Figure 4 provides an example of B-cos and other (post-hoc) explanations. It can be seen the B-cos explanation highlights important tokens well 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 provides clearer and more relevant attributions compared to the post-hoc explanations.

7 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 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 (cf. Appendix E). 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. Additionally, we report results for conventional fine-tuning (Conv.

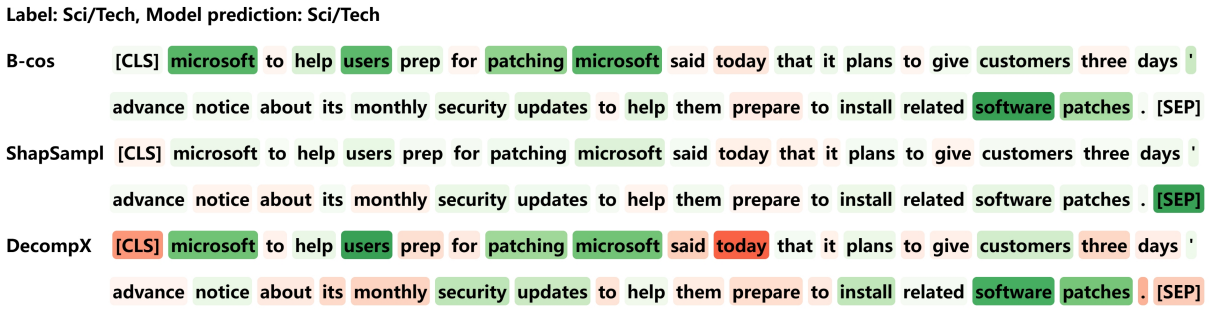

Figure 4: Examples of B-cos explanations (B-cos BERT) as well as ShapSampl and DecompX explanations (BERT) from AG News. Green (red) indicates the positive (negative) impact of tokens on the prediction. The B-cos explanation highlights only relevant tokens and is more interpretable to humans (cf. Appendix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

LM) and training a randomly initialized B-cos LM (B-cos from scratch). Experiments are conducted on IMDB, with results averaged over three runs.

Setup	Epochs	Acc (↑)	SeqPG (↑)	Steps (K)
Conv. LM	5	94.06	-	7.33
B-cos LM	5	89.85	70.41	3.67
B-cos from scratch	5	88.36	60.92	4.33
Task then B-cos	1+4	90.14	70.28	1+4.33
	2+3	90.33	70.36	3+3.33
	3+2	90.07	69.94	4+3
	4+1	88.19	70.36	5+1
	5+5*	90.33	69.65	6.67+3.33
B-cos then task	1+4	89.78	65.58	1+5.67
	2+3	89.81	66.01	3+4
	3+2	89.38	66.95	4+3
	4+1	87.42	67.9	6+1
	5+5*	90.38	71.16	7+3
	10+5*	91.08	75.06	15+3.67
	20+5*	91.75	76.66	31+6.33

Table 3: Different B-cosification setups. For two-phase methods, we report epoch distribution and convergence steps per phase. * marks additional training epochs.

Table 3 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 task performance 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 comparable to B-cos LM but requires more 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.

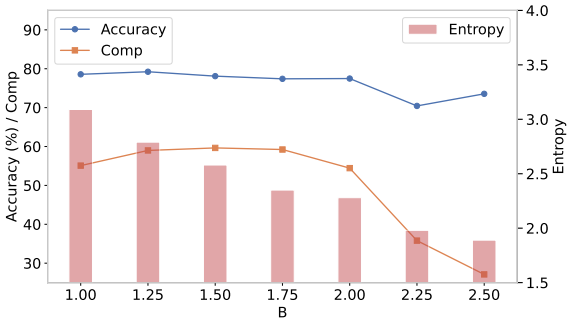

Figure 5: Varying B for B-cos BERT (HateXplain). Accuracy and Comp both peak around B=1.5, while explanation entropy negatively correlates with B.

8 Effects of B-cosification and B Values

For a deeper understanding of how B-cosification and 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 K.

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, peaking 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 L.

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 $|\cos(\mathbf{x}, \hat{\mathbf{W}})|^{\mathbf{B}-\mathbf{1}}$ 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 M 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 learn biases in the data. For this, we examine label bias and word-level spurious correlations using the HateXplain dataset, 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 prediction positive rate and lower 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 datasets.

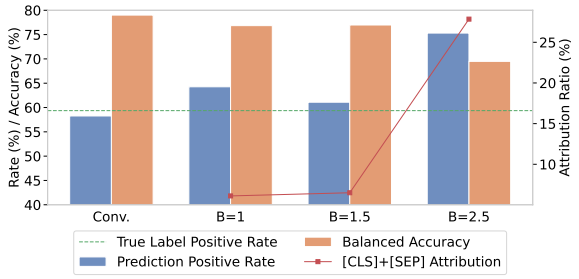

Figure 6: Comparison of conv. BERT and B-cos BERT with different B values. The attributions to [CLS] and [SEP] tokens (■) indicate that B-cos LMs with large B overfit to the non-semantic label distribution.

We also find that B-cosification—particularly with large B—amplifies the 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 N). However, the faithfulness and interpretability of B-cos explanations facilitate the identification 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.

9 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 4).

Method	Conv. BERT		B-cos BERT	
	Time (s)	Memory (GB)	Time (s)	Memory (GB)
ShapSampl	37.22	21.95	70.49	22.95
LIME	6.82	21.96	8.92	22.95
SIG	67.46	29.09	108.48	69.32
DecompX	0.76	48.38	-	-
B-cos	-	-	**0.08**	**2.78**

Table 4: Computational costs of generating explanations for 100 instances randomly sampled from IMDB (test) using an NVIDIA H100 GPU (batch size 1). We see that the B-cos explanations (**bold**) are at least 9x faster and require at most $\frac{1}{8}$ of VRAM.

10 Conclusion

In this work, we introduce B-cos LM, a 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.

11 Limitations

This study has certain limitations that should be acknowledged.

Firstly, 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.

Secondly, our study does not include a direct comparison with other methods designed to enhance model explainability, which may limit the scope of our findings. This omission is due to two reasons: (1) existing explainable models often provide only marginal improvements over post-hoc explanation methods (Brinner and Zarrieß, 2024), and (2) incorporating them into our study would require substantial computational resources, as many baseline explanation methods are computationally expensive.

Finally, although B-cos LMs can be applied to different model architectures and tasks, our experiments focus only on encoder-only models for sequence classification tasks. Extending our approach to other architectures and tasks remains an avenue for future work.

12 Ethical Considerations

As discussed in § 8, B-cos LMs can overfit to biases present in the training data. Although their more faithful and human interpretable explanations make biased predictions easier to detect, this does not eliminate the risk of unintended bias amplification. We encourage users to carefully assess potential biases in their specific use cases before deploying B-cos LMs and to incorporate bias mitigation strategies where necessary.

All models and datasets used in this work comply with their respective licenses. Their usage aligns with their intended purpose as specified by their creators.

The human study complies with all ethical research guidelines set by our institutes. All participants of the human evaluation study were master's or doctoral students with backgrounds in computer science or computational linguistics and were proficient in English. They were volunteers and were compensated with the standard hourly salary set by the university (at least 5% above minimum wage). Before participation, all participants were informed about the content and purpose of the study, the collected data and its usage. They were instructed on how they could access, modify, or delete their data post-study and provided their informed consent.

References

David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. 2019. Evaluating recurrent neural network explanations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 113–126, Florence, Italy. Association for Computational Linguistics.

Shreyash Arya, Sukrut Rao, Moritz Böhle, and Bernt Schiele. 2024. B-cosification: Transforming deep neural networks to be inherently linterpretable. In *Advances in Neural Information Processing Systems*, volume 37, pages 62756–62786. Curran Associates, Inc.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

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

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

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

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. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Carlo Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Marc Felix Brinner and Sina Zarrieß. 2024. Rationalizing transformer predictions via end-to-end differentiable self-training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11894–11907, Miami, Florida, USA. Association for Computational Linguistics.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Joseph Enguehard. 2023. Sequential integrated gradients: a simple but effective method for explaining language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7555–7565, Toronto, Canada. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

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. 2023. A framework for few-shot language model evaluation.

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

Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. 2021. Improving deep learning interpretability by saliency guided training. In *Advances in Neural Information Processing Systems*, volume 34, pages 26726–26739. Curran Associates, Inc.

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

Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. 2018. 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.

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

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

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

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

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

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

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 1–10. Curran Associates, Inc.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

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

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

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2023. DecompX: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2649–2664, Toronto, Canada. Association for Computational Linguistics.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2020. Training with adversaries to improve faithfulness of attention in neural machine translation. In *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*, pages 93–100, Suzhou, China. Association for Computational Linguistics.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2791–2802, Online. Association for Computational Linguistics.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 807–814. Omnipress.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Nina Poerner, Hinrich Schütze, and Benjamin Roth. 2018. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 340–350, Melbourne, Australia. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

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

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

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

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

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

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

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

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

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

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2015. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Suraj Srinivas and François Fleuret. 2019. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Shengjie Wang, Tianyi Zhou, and Jeff Bilmes. 2019. Bias also matters: Bias attribution for deep neural network explanation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6659–6667. PMLR.

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. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yiming Zheng, Serena Booth, Julie Shah, and Yilun Zhou. 2022. The irrationality of neural rationale models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 64–73, Seattle, U.S.A. Association for Computational Linguistics.

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 and 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. Additionally, 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 Ablation Study

To gain deeper insights into B-cosification, we conduct an ablation study to evaluate the effects of key design choices on model performance. Table 5 reports the effects of these modifications.

Consistent with § 8, B=1 results in worse task performance and lower explanation faithfulness. Using binary cross-entropy (BCE) loss instead of conventional cross-entropy loss has minimal impact on classification accuracy, but leads to better faithfulness results in perturbation-based evaluations. Additionally, architectural adaptations, including removing bias terms and eliminating activation functions in prediction heads, play a crucial role in improving both model performance and explainability in B-cos LMs. Besides, we encountered numerical instability when generating explanations without these architectural adaptations, as the dynamic linear weight for tanh ($\frac{\tanh(\mathbf{x})}{(\mathbf{x})}$) becomes unstable when $\mathbf{x}$ is close to 0.

Beyond ablating components in model design and training, we also examine different explanation methods across models. First, replacing dynamic linear weights $\mathbf{W}(\mathbf{x})$ with gradients for computing input contributions (equivalent to InputXGradient, Kindermans et al., 2016) results in less faithful explanations. Moreover, 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..[4]

[4]Extracting $\mathbf{W}(\mathbf{x})\mathbf{x}$ from conventional models follows the same approach as in B-cos LMs (cf. § D), except that in standard linear transformations, the dynamic linear weight is replaced by the fixed weight matrix $\mathbf{W}$.

	Acc (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)
Full system	78.64	59.66	-4.89	77.57
w/o alignment pressure (B=1)	78.07	57.19	-2.57	70.18
w/o BCE training	79.00	49.22	-7.91	79.21
w/o architectural adaptations	77.65	52.23	-3.80	74.30
w/o dynamic linear weights (IxG)	78.64	44.93	-0.60	53.57
$\mathbf{W}(\mathbf{x})\mathbf{x}$ from conv. model	80.77	44.92	2.80	70.20

Table 5: Ablation study of key designs in B-cos BERT model on HateXplain. Results are averaged over three runs.

D Dynamic Linear Representation of Model Components

Here we describe how each model component functions as a dynamic linear module in B-cos LMs.

B-cos Layers B-cos layers are designed as dynamic linear modules with a dynamic linear weight matrix $\mathbf{W}(\mathbf{x}) = |\cos(\mathbf{x}, \hat{\mathbf{W}})|^{\mathbf{B-1}} \otimes \hat{\mathbf{W}}$. 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 and Hinton, 2010) and GELU (Hendrycks and 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 a dynamic linear weight.

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

$$\begin{aligned}\text{Att}(\mathbf{X}; \mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{softmax}(\mathbf{X}^\mathbf{T}\mathbf{Q}^\mathbf{T}\mathbf{K}\mathbf{X})\mathbf{V}\mathbf{X} \\ &= \mathbf{A}(\mathbf{X})\mathbf{V}\mathbf{X} = \mathbf{W}(\mathbf{X})\mathbf{X} \qquad (4)\end{aligned}$$

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 $\mathbf{U}$:

$$\text{MSA}(\mathbf{X}) = \mathbf{U}[\mathbf{W}_1(\mathbf{X})\mathbf{X}, ..., \mathbf{W}_H(\mathbf{X})\mathbf{X})] \qquad (5)$$

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

E 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.

Post-hoc Explanation Baselines For IxG and ShapSampl, we use the Captum (Kokhlikyan et al., 2020) implementations.[5] We implement the Attention method ourselves, and LIME is sourced from the lit library[6]. For DecompX[7] and SIG[8], 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.

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 B-cos LMs on the same sets of examples constructed

[5] https://captum.ai/api/
[6] https://github.com/PAIR-code/lit
[7] https://github.com/mohsenfayyaz/DecompX
[8] https://github.com/josephenguehard/time_interpret

based on the predictions of the corresponding conventional models.

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.

B-cos Pre-training For B-cos pre-training in § 7, we further pre-train the model on the Wikipedia dataset[9] using masked language modeling loss with a learning rate of 1e-4 and a 15% masking ratio.

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.

F 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 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.

G Task Performance of Other B-cos LMs

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.

H Faithfulness Evaluation of Other B-cos LMs

Tables 6 and 7 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.

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

I 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.

> **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 50 examples where NLP models perform classification tasks and provide explanations for their predictions.
>
> - The first 25 examples come from a hate speech detection task, where the model predicts whether a text is toxic or not toxic.
> - The last 25 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

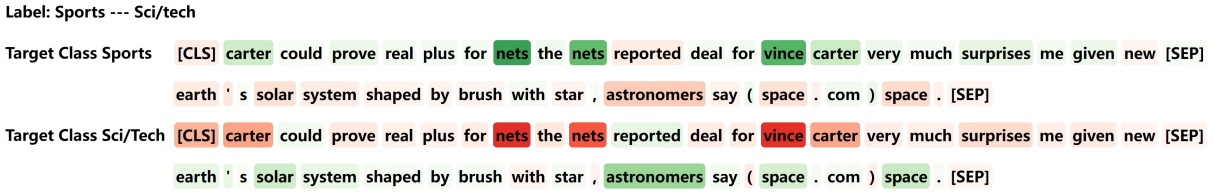

Figure 7: An example of SeqPG from AG News (using B-cos BERT). Green (red) indicates the positive (negative) impact of tokens on the prediction. The example consists of two sequences with different labels (Sports and Sci/tech), separated by the [SEP] token after the first sequence. Explanations are generated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

Model	Method	AG News			IMDB			HateXplain		
		Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)
(a) 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) B-cos DistilBERT	Attention	28.47	3.05	50	31.36	4.15	50	37.33	6.49	50
	IxG	22.33	9.09	58.63	51.02	-1.44	53.76	41.62	0.29	56.03
	SIG	14.73	5.62	53.09	39.75	-0.11	64.18	28.68	7.27	60.75
	DecompX	-	-	-	-	-	-	-	-	-
	ShapSampl	31.78	1.77	62.60	64.65	-2.42	56.89	34.64	4.56	55.8
	LIME	58.25	0.31	77.65	69.96	-0.43	61.08	44.66	1.66	59.27
(c) B-cos DistilBERT	B-cos	**61.93**	**-1.01**	**86.78**	**75.73**	**-2.57**	**71.95**	**57.2**	**-4.49**	**74.89**

Table 6: Faithfulness evaluation for conventionally fine-tuned DistilBERT and B-cos DistilBERT across three datasets. The best results are in **bold**. We find that B-cos explanations are consistently more faithful than post-hoc explanations from both models.

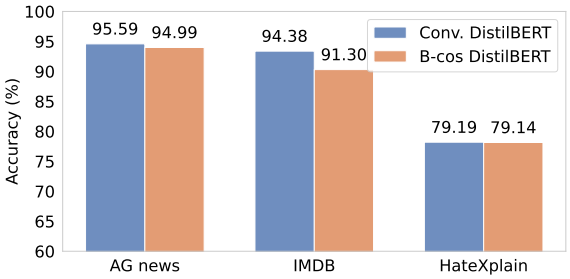

Figure 8: Mean accuracy of conventionally fine-tuned and B-cos DistilBERT models averaged over three runs. B-cos models perform comparably to conventional models on most tasks.

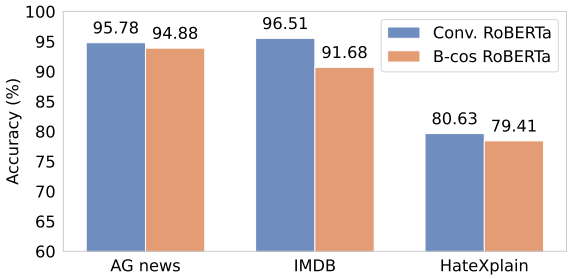

Figure 9: Mean accuracy of conventionally fine-tuned and B-cos RoBERTa models averaged over three runs. B-cos models perform comparably to conventional models on most tasks.

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.

Model	Method	AG News			IMDB			HateXplain		
		Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)	Comp (↑)	Suff (↓)	SeqPG (↑)
(a) 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
(b) B-cos RoBERTa	Attention	16.07	6.83	50	29.83	2.85	50	27.35	8.39	50
	IxG	22.25	2.39	56.15	67.2	-2.26	56.95	40.69	-1.11	58.59
	SIG	44.35	-0.95	51.70	74.70	**-2.39**	58.03	51.20	**-5.80**	57.62
	DecompX	-	-	-	-	-	-	-	-	-
	ShapSampl	55.26	-0.92	72.65	74.3	**-2.39**	62.74	**51.54**	-5.64	70.58
	LIME	23.01	2.46	63.11	37.14	0.78	53.15	29.86	5.74	63.61
(c) B-cos RoBERTa	B-cos	**62.47**	**-1.18**	86.63	**75.15**	**-2.39**	**75.83**	51.33	-5.18	**74.01**

Table 7: Faithfulness evaluation for conventionally fine-tuned RoBERTa and B-cos RoBERTa across three datasets. The best results are in **bold**. We find that B-cos explanations are consistently more faithful than post-hoc explanations from both models.

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.

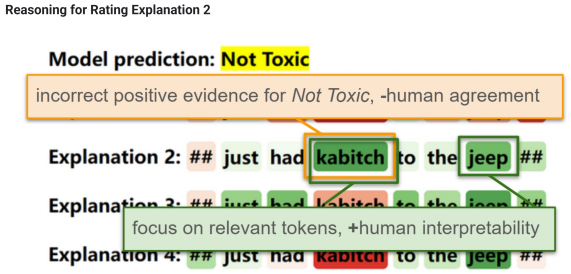

Figure 10: An example shown to participants that demonstrates how to rate explanations.

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.

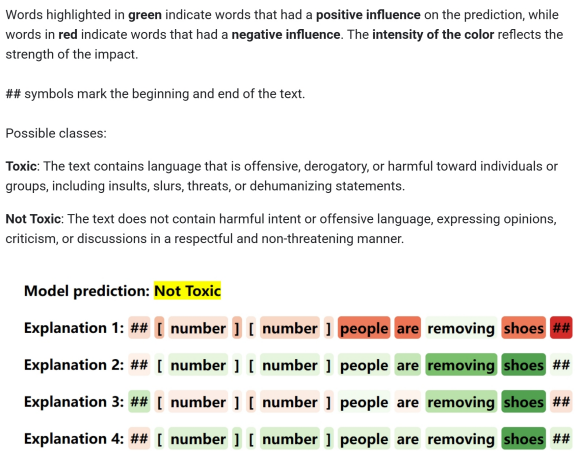

Figure 11: An examples of a model prediction and its explanations presented to participants.

J 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 § 6, B-cos LMs provide more human interpretable explanations.

K 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.

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

L Effects of B on Other Metrics

Table 8 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 8: 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.

M 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 § 8.

N 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 §8.

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.

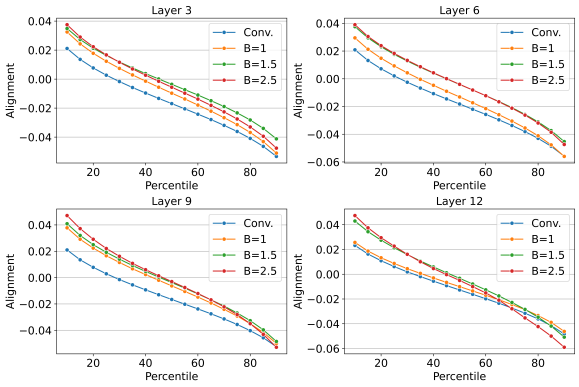

Figure 13: Percentiles of input-weight alignment in B-cos modules across selected layers of conventional and B-cos BERT models with different B values (HateXplain).

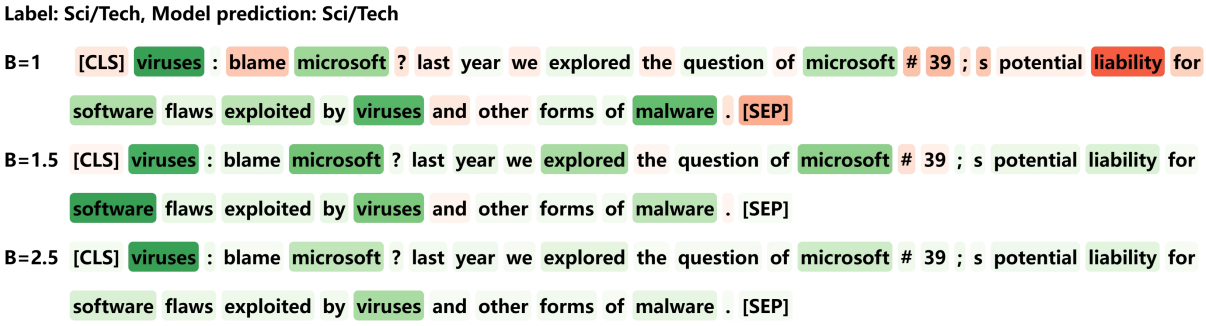

Figure 14: B-cos explanations (B-cos BERT) on AG News with different B values. Green (red) indicates the positive (negative) impact of tokens on the prediction. As B increases, B-cos LMs produce sparser explanations, with fewer tokens receiving significant attribution scores.

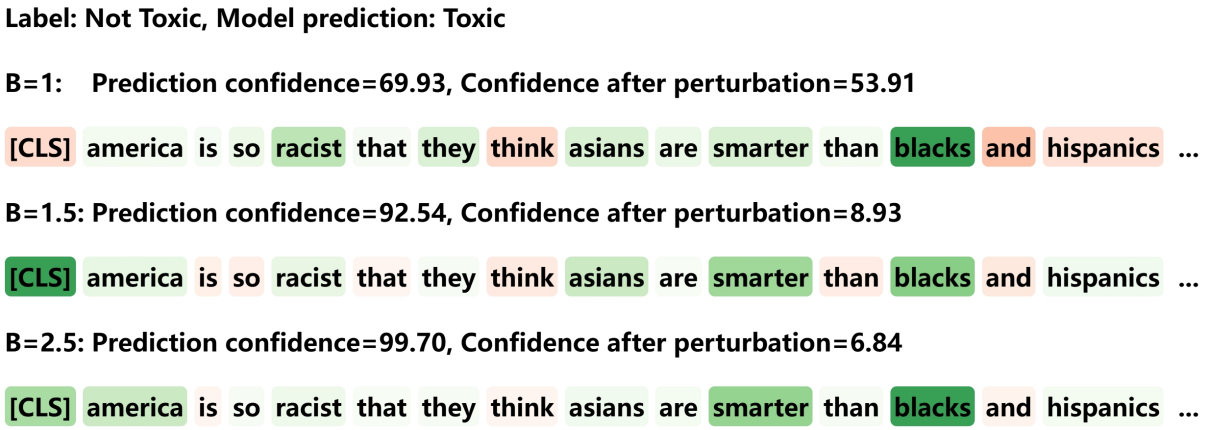

Figure 15: Example of how larger B values lead B-cos LMs to learn word-level spurious correlations. Green (red) indicates the positive (negative) impact of tokens on the prediction. Higher alignment pressure increases the reliance of B-cos LMs on spurious correlations in the data. In this example, perturbation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