Explanation Regularisation through the Lens of Attributions

Pedro Ferreira [1] **Ivan Titov** [1,2] **Wilker Aziz** [1]
[1] University of Amsterdam [2] University of Edinburgh
{ p.m.ferreira, w.aziz }@uva.nl ititov@inf.ed.ac.uk

Abstract

Explanation regularisation (ER) has been introduced as a way to guide text classifiers to form their predictions relying on input tokens that humans consider plausible. This is achieved by introducing an auxiliary explanation loss that measures how well the output of an input attribution technique for the model agrees with human-annotated rationales. The guidance appears to benefit performance in out-of-domain (OOD) settings, presumably due to an increased reliance on 'plausible' tokens. However, previous work has under-explored the impact of guidance on that reliance, particularly when reliance is measured using attribution techniques different from those used to guide the model. In this work, we seek to close this gap, and also explore the relationship between reliance on plausible features and OOD performance. We find that the connection between ER and the ability of a classifier to rely on plausible features has been overstated and that a stronger reliance on plausible tokens does not seem to be the cause for OOD improvements. [1]

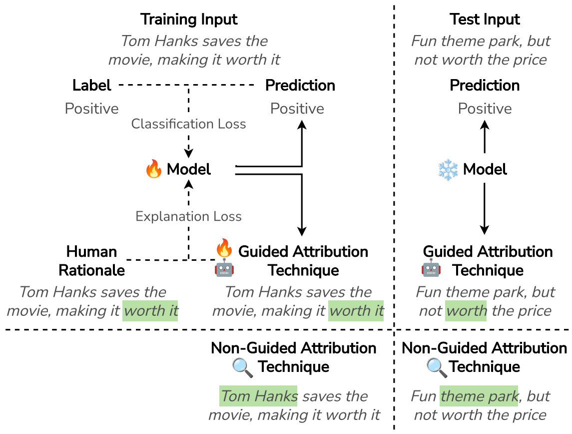


Figure 1: ER (top-left) minimises a *classification* and an *explanation* loss. The latter uses an input attribution technique to obtain a machine rationale for the prediction, and penalises differences between that and the human rationale. Guiding the model to rely in its predictions on plausible tokens is expected to help the classifier at test time (top-right), when no human rationale is available. Bottom: even though the attribution technique used for guidance shows human-like rationales, the model may in fact rely on different (non-plausible) tokens, as other attribution techniques might reveal.

1 Introduction

In explanation regularisation (ER; Ross et al., 2017; Ghaeini et al., 2019; Liu and Avci, 2019; Ismail et al., 2021; Joshi et al., 2022; Pruthi et al., 2022; Stacey et al., 2022, *i.a.*), a text classifier is encouraged to inform its predictions by tokens included in a human rationale for the label. This is achieved with an auxiliary 'explanation loss' that penalises differences between the output of a *guided attribution technique* (*e.g.*, a relevance map based on top-layer attention) and the annotated *human rationale* (see Figure 1). Compared to their counterparts trained without rationales, these ER models have been reported to improve classification, including for out-of-domain (OOD) inputs (Joshi et al., 2022; Stacey et al., 2022; Madani and Minervini, 2023), with this improved robustness being ascribed to increased reliance of ER models on *plausible tokens* (*i.e.*, those in the human rationale for the label).

While ER is assumed to encourage classifiers to rely more on plausible tokens, this has not been carefully verified in previous work. In particular, prior analyses have focused on guided attributions—*i.e.*, those *explicitly supervised* to resemble human rationales through the explanation loss—finding them to be better aligned with human rationales (Mathew et al., 2021; Stacey et al., 2022; Madani and Minervini, 2023, *i.a.*). Although impact on guided attributions is a necessary condition, this impact is not sufficient to demonstrate that the classifier relies more on plausible tokens. We argue

[1] Source code available at https://github.com/PedroMLF/ER_through_the_lens_of_attributions.

Proceedings of the 31st International Conference on Computational Linguistics, pages 6530–6551
January 19–24, 2025. ©2025 Association for Computational Linguistics

that, to confirm an increased reliance, it is also necessary to examine the impact of guidance on attributions obtained through other *non-guided* attribution techniques (see Figure 1). If these are unaffected, it would suggest that the model is 'hacking' (Skalse et al., 2022) the explanation loss rather than genuinely increasing its reliance on plausible tokens.

In this work, we analyse *(i)* ER's ability to make classifiers rely on plausible tokens; and *(ii)* the relationship between plausibility and robustness to OOD conditions. We start by studying ER with jointly optimised losses (§5.1). We find that – unlike the guided attribution techniques – there is little to no evidence that attributions by any other non-guided attributions are affected by ER, suggesting the loss 'hacking'. In fact, we find that meaningful impact on input attributions is only achievable by ensuring a low explanation loss, for example via constrained optimisation (see §5.2). However, this is effective only for 'global' guided attribution techniques (which capture the information 'flow' across all or most components of the model), and, crucially, at the expense of classification performance.

Contributions. We show that the connection between ER and the ability of a classifier to rely on plausible tokens has been overstated. In particular, we find that the choice of guided attribution technique is critical for this assumption: local attribution techniques, such as top-layer attention, can be 'hacked', minimising the explanation loss without much evidence of increased reliance on plausible tokens. Moreover, we find that OOD classification performance degrades for ER models that (over-)rely on plausible tokens, hinting at the need of finding better design choices for ER.

2 Related Work

Learning from explanations. Human explanations (*e.g.*, token-level rationales, as in Figure 1) have been used to improve text classifiers (Hartmann and Sonntag, 2022; Hase and Bansal, 2022). Examples include *select-predict* 'pipelines', where an explanation output by a first module serves as input to a classifier (Camburu et al., 2018), and *multi-task* settings, where a classifier and a rationale extractor are jointly trained (Carton et al., 2022; Chan et al., 2022). In contrast, we are interested in *explanation regularisation*, where, rather than having additional model components trained to perform rationale extraction, a model is trained to align *attributions* provided by a differentiable attribution technique with human rationales (Ross et al., 2017; Ghaeini et al., 2019; Liu and Avci, 2019; Rieger et al., 2020; Mathew et al., 2021; Pruthi et al., 2022; Stacey et al., 2022, *i.a.*). Most work on ER focuses on in-domain data, however, there is evidence that ER can improve robustness to OOD data, for text (Rieger et al., 2020; Joshi et al., 2022; Stacey et al., 2022), and image classification (Chefer et al., 2022; Rao et al., 2023). Other works identify difficulties in using explanations to learn better classifiers and propose methods to improve the compromise between classification performance and model explainability. These works differ from ours by: *(i)* not using human-annotated explanations (Plumb et al., 2020); or *(ii)* by departing significantly from the ER formulation, for example, by using a multi-task setup with a separate explainer (Carton et al., 2022) or by not employing a trainable explainer and instead incorporating human explanations through contrastive learning (Resck et al., 2024). We, instead, focus on ER and explore its impact on the plausibility of input attributions and how that, in turn, impacts classification performance both in- and out-of-domain.

Attribution techniques. In *vector-based* techniques, model components are directly used to obtain input attributions. The simplest approach is to use attention weights (Clark et al., 2019, *i.a.*), despite conflicting evidence on its usefulness for this purpose (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Pruthi et al., 2020). Later works (Kobayashi et al., 2020, 2021; Ferrando et al., 2022; Modarressi et al., 2022) incorporate information about the magnitude of input vectors and the influence of other components of the Transformer layer (Vaswani et al., 2017). These are examples of *local* techniques, where attributions are based on the dynamics of a single layer.

It is also possible to obtain a *global* analysis of the model, by including the dynamics of its multiple layers. One example is *attention-rollout* (Abnar and Zuidema, 2020), which recursively aggregates attention weights across layers. Other examples of global vector-based techniques are GlobEnc (Modarressi et al., 2022), ALTI (Ferrando et al., 2022), and DecompX (Modarressi et al., 2023). Global attributions can also be obtained from *gradient-based* approaches (Kindermans et al., 2016; Shrikumar et al., 2017; Sundararajan et al., 2017, *i.a.*), or a mix of attention and gradients (Chefer et al., 2021; Qiang et al., 2022).

3 Explanation Regularisation

In ER (Joshi et al., 2022; Stacey et al., 2022; Pruthi et al., 2022, *i.a.*), we assume the availability of training data $\mathcal{D} = \{(x_i, y_i, e_i)_{i=1}^N\}$, for each data point, x_i is the input text, y_i is the target label, and e_i is a human rationale (specifically, a subset of tokens of the input regarded as a justification for the label). Human rationales can be obtained from annotators (Socher et al., 2013; Camburu et al., 2018; Rajani et al., 2019; Mathew et al., 2021) or through heuristics (Liu and Avci, 2019; Rieger et al., 2020), and are only required during training.

An ER classifier minimises a joint loss

$$\mathcal{L}_{\text{cls}}(\theta) + \lambda\, \mathcal{L}_{\text{expl}}(\theta), \tag{1}$$

where $\mathcal{L}_{\text{cls}}$ is the regular classification loss and $\mathcal{L}_{\text{expl}}$ is the *explanation loss*, with a hyperparameter $\lambda \in \mathbb{R}_{\geq 0}$ controlling the contribution of the latter. The expectation is that $\mathcal{L}_{\text{expl}}$ leads the model to rely in its decisions on more plausible tokens.

The first term is the cross-entropy loss $\mathcal{L}_{\text{cls}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}_{\text{ce}}(C_\theta(x), y)]$, where C_θ is a trainable classifier. The second term, $\mathcal{L}_{\text{expl}}(\theta) = \mathbb{E}_{(x,y,e)\sim\mathcal{D}}[\mathcal{S}(E(C_\theta, x), e)]$, employs an attribution technique in order to obtain a relevance map $E(C_\theta, x)$ for the classifier's output and penalises the classifier by this map's deviation from the human rationale, as assessed by a function $\mathcal{S}$ such as mean absolute error or KL divergence (Joshi et al., 2022). We refer to E as *the guided attribution technique*. Besides being differentiable, E is required to be memory- and time-efficient. Examples of techniques used in ER include gradient-based (Ross et al., 2017; Ghaeini et al., 2019; Chefer et al., 2022), vector-based (Mathew et al., 2021; Pruthi et al., 2022; Stacey et al., 2022; Fernandes et al., 2022), and perturbation-based approaches (Ying et al., 2022). We follow Joshi et al. (2022) and Stacey et al. (2022) and use top-layer attention and InputXGradient (Shrikumar et al., 2017). In addition, we also experiment with attention-rollout (Abnar and Zuidema, 2020).

4 Experimental Setting

Data. We use SST-2 (Socher et al., 2013) as training and in-domain data, following the heuristic proposed by Carton et al. (2020) to obtain instance-level rationales. For OOD data with rationales we use Movies (Zaidan and Eisner, 2008; DeYoung et al., 2020), and annotate Yelp-50, a subset of Yelp, as described in Appendix G.1. For OOD without rationales annotations we use Amazon Reviews ('Movies and TV' split) (Hou et al., 2024), IMBD (Maas et al., 2011), and Yelp (Zhang et al., 2015). For more details on data refer to Appendix G.

Model. All experiments use HuggingFace Transformers' (Wolf et al., 2020) BigBird-roberta-base model as the pre-trained contextual embedding encoder model (Zaheer et al., 2020). This choice follows previous works (Chan et al., 2022; Joshi et al., 2022; Madani and Minervini, 2023). For more training details refer to Appendix H.

Scoring Function. We follow Joshi et al. (2022) and use mean absolute error for $\mathcal{S}$ in $\mathcal{L}_{\text{expl}}(\theta)$.

Target Annotations. All datasets with rationales are annotated at the instance level with binary vectors, whose dimensionality equals the number of tokens. A value of 1 indicates relevant, highlighted tokens. In order to obtain a target for the scoring function $\mathcal{S}$, the instance-level vector is normalized to sum to one by dividing each element by the total number of highlighted tokens in the instance when using top-layer attention and attention-rollout (Joshi et al., 2022; Stacey et al., 2022). For InputXGradient we use the original binary values.

Attribution Techniques. We make use of one local technique, namely, top-layer attention (Att), and 5 global techniques, namely, attention rollout (AttR; Abnar and Zuidema, 2020), InputXGradient (IxG; Shrikumar et al., 2017), ALTI (Ferrando et al., 2022), and DecompX (Modarressi et al., 2023) with (DX-C) and without (DX) the classification head. [2] For guidance we use Att, AttR or IxG, as those are memory- and time-efficient enough to be used during training. The remainder are used only for analysis. All techniques are further described in Appendix A.

Plausibility Metrics. To assess how well input attributions reproduce the human annotated rationales, we use three metrics introduced in Fomicheva et al. (2021), and described in Appendix D: AUC Score, Average Precision; and Recall@k.

Constrained Optimisation. To study classifiers whose guided attributions are as plausible as they can be (through the lens of the explanation loss), we can reimagine ER as a constrained optimisation

[2] For Att and AttR, we average attention weights across heads. For IxG we use the Captum package (Kokhlikyan et al., 2020). For DecompX and ALTI we adapt the original code.

problem: minimise classification loss subject to a bound b on the explanation loss,

$$\min_{\theta} \quad \mathcal{L}_{\text{cls}}(\theta) \quad \text{s.t.} \quad \mathcal{L}_{\text{expl}}(\theta) \leq b \;, \tag{2}$$

with b set to a value close to $\min_\theta\ \mathcal{L}_{\text{expl}}(\theta)$ – a minimiser of the explanation loss, taken in isolation. We approach this via Lagrangian relaxation (Boyd and Vandenberghe, 2004), with implementation choices detailed in Appendix H.

5 Results

ER	Attr.	F1	Guided	Non-Guided
Joint (§5.1)	Local	✓	✓	∅
	Global	✓	∅	∅
Constr (§5.2)	Local	✓	✓	∅
	Global	×	✓	✓

Table 1: Summary of the effect (✓ positive, × negative, or ∅ nil) of local or global guidance, across Joint and Constrained setups, in terms of OOD classification (F1) and impact on guided and non-guided attributions.

Table 1 presents a brief overview of our observations, which are discussed in detail in the following subsections. In §5.1, we find that despite positively impacting the OOD generalisation of the classification task, joint ER has limited impact on input attributions. In fact, only local guidance is able to better predict plausible rationales, and that effect is only observed for the guided technique itself. With global guidance, we find that the explanation loss is under-optimised. To understand what is happening, in §5.2 we reinterpret ER as constrained optimisation, where the explanation loss is constrained to be low. At this point, we find it to be possible to affect input attributions meaningfully, but only when guiding a global attribution technique and, crucially, at the expense of classification performance. In a nutshell, local guidance is being 'hacked': attributions are modified locally as to optimise the explanation loss without affecting the features used for classification (intuitively, the model 'hides' non-plausible computations where the local attribution technique cannot 'see', such as in lower layers). Global guidance, on the other hand, seems much harder to 'hack': optimising it well requires modifying computations performed in different layers across the model, more strongly restricting the features used for classification to rationale tokens. This, however, tends to worsen classification performance. Finally, we find in §5.3 that the disconnect between (over-)reliance on plausible tokens and OOD classification performance prevents us from systematically using features available for model selection to find models that perform best OOD.

5.1 Joint Optimisation

We start by studying the most common ER setup (§3), where the two losses are jointly optimised.

Joint ER improves OOD classification. As observed in Table 2, joint ER improves OOD average classification performance. This is more noticeable for guidance with local attention (ER+ATT), with guidance that uses global attribution techniques (ER+ATTR and ER+IXG) showing smaller improvements. If we consider variance across runs, and visualise the distribution of results across seeds (Fig. 2), it is noticeable how the effect on the average improvement is exacerbated by runs that perform poorly. In fact, this seems to indicate that one of the merits of ER is to converge less often to models that generalise poorly to OOD conditions.

Only local guidance improves rationale extraction performance. In addition to classification, ER models are trained to align guided attributions to human rationales. Hence, compared to the baseline, we expect an impact on the plausibility of the corresponding guided attributions. This effect can be observed by comparing the *highlighted* cells in Table 3 with the corresponding baseline. Our expectation is met *only* with local guidance, with average AUC plausibility increasing from 35.6 (baseline) to 78.6 (w/ ATT) in-domain and similar, but more modest, improvements OOD (Table 4).

The lack of impact of global guidance on the plausibility of the corresponding guided attributions might come as a surprise. However, we need to consider that: *(i)* this class of techniques incorporates more of the model's components when computing attributions, potentially making it more difficult for the model to 'hide' implausible computations from the explanation loss; and *(ii)* joint ER balances both classification and explanation losses with a preference (via model selection) for classification, meaning that to maintain classification performance the model might require 'under-optimising' the explanation loss. Both possibilities will be further studied in this section.

ER+ATT fails to impact non-guided attribution techniques. The previous results illustrate how attention guidance (ER+ATT) is able to simulta-

		SST Dev	*SST Test*	**Movies**	**Yelp**	**IMDB**	**Amazon-M-TV**
	BASELINE	92.95 ± 0.57	94.36 ± 0.48	91.01 ± 4.38	94.41 ± 0.99	91.20 ± 1.63	83.89 ± 0.89
J §5.1	ATTENTION	93.56 ± 0.50	94.33 ± 0.72	93.10 ± 2.44	95.09 ± 0.44	91.87 ± 1.10	84.10 ± 0.56
	+ ROLLOUT	93.19 ± 0.68	94.09 ± 0.82	92.72 ± 2.71	94.77 ± 0.66	91.88 ± 0.99	84.15 ± 0.47
	IXG	92.91 ± 0.58	94.13 ± 0.56	91.11 ± 3.37	94.72 ± 0.54	91.30 ± 1.44	84.29 ± 0.92
C §5.2	ATTENTION	93.41 ± 0.57	94.45 ± 0.49	90.59 ± 4.80	94.52 ± 1.30	90.65 ± 2.17	83.38 ± 1.11
	+ ROLLOUT	90.78 ± 0.90	91.11 ± 0.77	89.56 ± 2.67	92.40 ± 0.78	89.19 ± 0.85	81.44 ± 0.81
	IXG	90.64 ± 0.90	91.59 ± 1.46	75.80 ± 15.1	88.99 ± 6.17	83.35 ± 7.58	79.95 ± 2.57

Table 2: F1-Macro (↑) and standard deviation for *in-domain (SST)* and **OOD data**. ATTENTION corresponds to an ER (J)oint or (C)onstrained model that uses attention as the guided attribution technique, +ROLLOUT to a model that uses attention-rollout, and IXG to a model that uses INPUTXGRADIENT. Results are averages of 15 seeds.

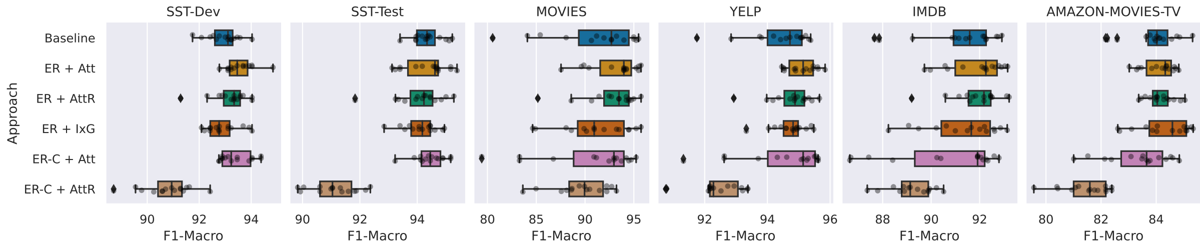


Figure 2: F1-Macro scores (↑). ER + ATT uses attention as the guided attribution technique, ER + ATTR uses attention-rollout, and ER + IXG uses INPUTXGRADIENT. The C prefix indicates a constrained model. Results correspond to 15 seeds. ER-C + IXG is not shown for improved clarity and can be seen in Appendix Figure 11.

neously improve OOD classification and rationale extraction (when assessed with the corresponding guided attribution technique). It is tempting to connect these two observations and conclude that ATT guidance results in ER models that better classify *while* relying more on plausible tokens. However, we find that not to be the case. Firstly, we correlate *input attributions* (obtained by a given technique) across training conditions (*e.g.*, ALTI baseline attributions vs. ALTI ER+ATT attributions). We find that – aside from the guided attribution technique – correlation coefficients are mostly unaffected by ER training, both in- and out-of-domain (see first row of Figure 3, for OOD). We find similar evidence for all non-guided techniques (Appendix Fig. 9). Secondly, we compare the AUC plausibility scores of the non-guided techniques—any non-highlighted cell in Tables 3 and 4—with the corresponding baselines. Here, we observe a lack of impact across joint ER models. For example, for ER+ATT we find the impact to be small (59.0 vs 64.5 for ALTI) to none (58.5 to 58.4 for DX-C). We find similar evidence using average precision and recall@k (Appendix Table 9). In fact, if we look at AUC plausibility scores across layers (Fig. 4) we can observe that ER+ATT only impacts the plausibility of attributions at the top-layer. This result confirms our suspicions: local guidance is

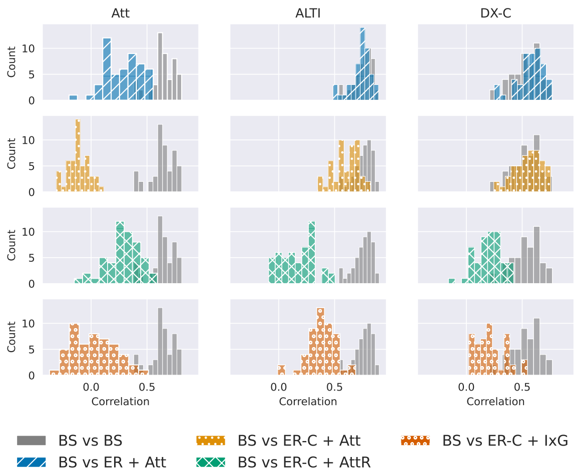


Figure 3: **Yelp-50** (OOD) Kendall Rank correlations between attribution techniques for different approaches. Baseline (BS) vs Baseline serves as ground-truth for the expected correlations agreement due to seed variability.

able to 'hide' implausible computations at lower layers, thus hiding them from $\mathcal{L}_{\text{expl}}$.

$\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{expl}}$ exhibit no synergy. We observed before how ER+ATTR and ER+IXG failed to impact input attributions, including for the respective guided attribution technique. To understand why this happened, we train ER models exclusively with the explanation loss, $\mathcal{L} = \mathcal{L}_{\text{expl}}$, and obtain a

	SST-Dev					
	Att	AttR	IxG	ALTI	DX-C	DX
Baseline	35.6	27.9	50.7	59.0	58.5	57.5
Joint ER (§5.1)						
w/ Att	*78.6*	27.9	50.6	64.5	58.4	57.8
w/ AttR	37.0	*28.7*	50.8	60.2	58.6	57.4
w/ IxG	44.6	27.6	*53.8*	61.7	58.0	57.6
Constrained ER (§5.2)						
w/ Att	*87.6*	28.2	51.5	68.3	57.6	56.8
w/ AttR	63.2	*89.1*	67.6	87.9	76.9	81.3
w/ IxG	63.9	42.9	*81.5*	74.8	61.3	64.8
$\mathcal{L} = \mathcal{L}_{\text{expl}}$ (§5.1)						
$\mathcal{L}_{\text{expl}}$ (A)	*89.0*	29.6	53.2	78.2	52.0	53.0
$\mathcal{L}_{\text{expl}}$ (R)	81.1	*89.4*	72.9	88.2	77.1	83.3
$\mathcal{L}_{\text{expl}}$ (I)	69.2	57.4	*84.3*	81.0	65.2	66.1

Table 3: Average **in-domain** AUC plausibility scores (↑). $\mathcal{L}_{\text{expl}}$ (A/R/I) are trained with $\mathcal{L} = \mathcal{L}_{\text{expl}}$, for Att, AttR, and IxG respectively. *Highlighted values* correspond to the ER guided attribution technique.

	Movies			Yelp-50		
	Att	AttR	IxG	Att	AttR	IxG
Baseline	70.0	49.6	58.1	51.9	37.1	51.5
Joint ER (§5.1)						
w/ Att	*74.6*	49.1	58.1	*69.7*	36.4	50.7
w/ AttR	70.1	*50.2*	58.0	52.5	*37.6*	51.7
w/ IxG	69.9	49.0	*59.5*	56.5	37.3	*53.6*
Constrained ER (§5.2)						
w/ Att	*68.8*	48.1	56.0	*76.0*	37.4	51.3
w/ AttR	71.1	*62.8*	59.3	67.2	*69.4*	58.8
w/ IxG	56.7	56.1	*61.6*	60.8	50.5	*69.8*
$\mathcal{L} = \mathcal{L}_{\text{expl}}$ (§5.1)						
$\mathcal{L}_{\text{expl}}$ (A)	*61.0*	48.2	48.0	*70.1*	38.5	47.1
$\mathcal{L}_{\text{expl}}$ (R)	59.4	*62.5*	55.2	62.6	*68.6*	56.8
$\mathcal{L}_{\text{expl}}$ (I)	51.8	56.9	*59.8*	55.2	58.3	*68.6*

Table 4: Average **out-of-domain** AUC plausibility scores (↑). $\mathcal{L}_{\text{expl}}$ (A/R/I) are trained with $\mathcal{L} = \mathcal{L}_{\text{expl}}$, for Att, AttR, and IxG respectively. *Highlighted values* correspond to the ER guided attribution technique.

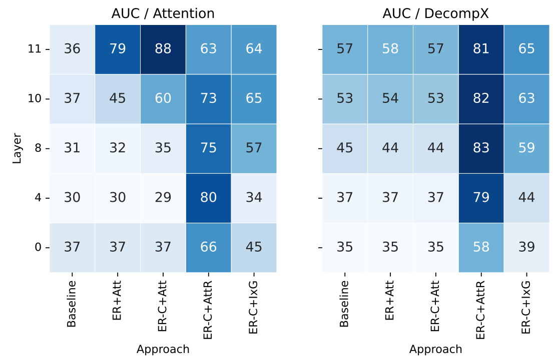


Figure 4: **SST-Dev** average AUC plausibility score per layer (↑) with Attention and DecompX.

E_c	Objective	$\mathcal{L}_{\text{expl}}^{(\text{dev})}$ (↓)	$\text{AUC}^{(\text{dev})}$ (↑)
Att	$\mathcal{L}_{\text{expl}}$	0.030	89.0
	(J) ER	0.040	78.6
	(C) ER	0.034	87.6
AttR	$\mathcal{L}_{\text{expl}}$	0.032	89.4
	(J) ER	0.064	28.7
	(C) ER	0.033	89.1
IxG	$\mathcal{L}_{\text{expl}}$	0.349	84.3
	(J) ER	0.432	53.8
	(C) ER	0.358	81.5

Table 5: **SST-Dev** explanation loss and AUC values for models guided with Att, AttR, or IxG, as a function of their training objective (explanation loss alone, vs. (J)oint ER and (C)onstrained ER).

'lower-bound' for its value. As seen in Table 5, all joint ER approaches exhibit a gap to the explanation loss 'lower-bound' value, particularly AttR and IxG, with this observation extending to AUC — *e.g.*, for $\mathcal{L}_{\text{expl}}$ (A) we obtain a AUC value of 89.0 (vs the 78.6 obtained with joint ER+Att), and for $\mathcal{L}_{\text{expl}}$ (R) we obtain a value of 89.4 (vs the 28.7 obtained with joint ER+AttR). These results highlight one possible difficulty of the joint ER setup: promoting plausible attributions via ER with global guidance might incur a negative cost in classification performance, hence joint optimisation (whose hyperparameters are selected based on classification loss) will under-optimise the auxiliary task. In order to confirm this behavior we run the joint ER setup with increasing values of λ. The results for ER+Att and ER+AttR can be seen in Fig. 5. We observe two overall trends. Given a high enough λ, the explanation loss converges to the obtained 'lower-bound'. For ER+Att, even smaller values of λ impact the explanation loss, whereas for ER+AttR larger values are required. For both strategies alike, as the explanation loss values decreases, the cross entropy loss increases (we verify that all models still converge on the classification loss). However, the ER+Att model seems to be more robust to this trade-off—we observe a smaller increase in CE loss as the explanation loss decreases when compared to ER+AttR. We observe a similar effect for ER+IxG in Appendix Fig. 12.

Although this is, perhaps, not a necessarily

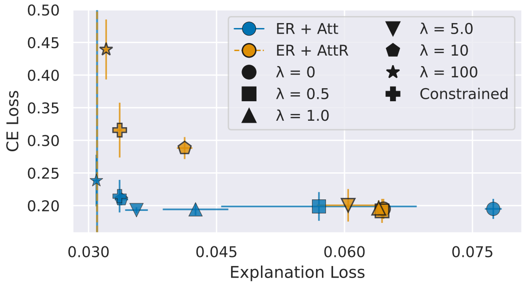


Figure 5: **SST-Dev** $\mathcal{L}_{\text{ce}}$ vs. $\mathcal{L}_{\text{expl}}$. The vertical lines show the validation explanation-loss bounds. Each point is the average of 5 runs, the error bars show standard deviation for each loss. λ ranges from 0 (No-ER) to 100.

surprising finding (Carton et al. (2022), for example, observe a similar tradeoff, though in a multi-tasking setting), it challenges a key assumption of ER as a means to improve OOD generalisation: that the explanation loss should help inform the classifier, improving OOD performance. In fact, we observe that the joint ER setup will struggle to accommodate both losses, particularly when using a guided attribution technique that is more difficult to 'hack' in order to become apparently better at predicting human rationales, such as attention-rollout, where the explanation loss might be lowered, but far from optimally.

5.2 Constrained Optimisation

In §5.1, we observed how joint ER under-optimises the explanation loss and how this effect varies in function of the guided attribution technique being local or global. Moreover, Table 3 shows that optimising exclusively the explanation loss, $\mathcal{L} = \mathcal{L}_{\text{expl}}$, can lead to AUC plausibility improvements—on the guided attribution technique alone for local guidance, and across the board for global guidance. This serves as motivation to analyse classifiers optimised for the classification task, while subject to achieving the aforementioned explanation loss 'lower-bound' via constrained optimisation.

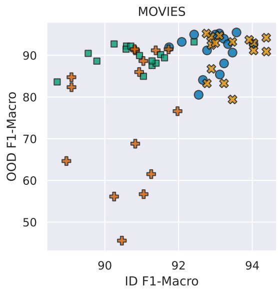
Constrained global guidance is necessary to impact non-guided attribution techniques. We start by noting how constrained ER addresses an earlier result: both global guided techniques are now impacting the corresponding guided attribution technique. This is true for both the AUC plausibility scores (in Table 3) and correlation coefficients (in Appendix Figure 9). However, this is so by design, and a more interesting observation stems from inspecting the non-guided attribution techniques — while constrained ER-C+ATT still mostly fails to impact other attribution techniques, constrained ER-C+ATTR and ER-C+IXG impact the outcome of all assessed attribution techniques. We can observe this along three 'dimensions'. First, global techniques are able to clearly impact the plausibility of all attribution techniques (Table 3). For instance, ER-C+ATTR leads to large improvements in DecompX, where the average AUC plausibility score is improved from 57.5 (in baseline) to 81.3, with the value being 56.8 for ER-C+ATT. Second, there is a clear impact in the correlation of attributions across approaches (Appendix Fig. 9a), contrarily to what is observed for local ER, including ER-C+ATT. Finally, AUC plausibility scores per layer (Figure 4) also show how the global technique is able to impact not only attributions at the top-layer, but across the whole model, making them more aligned with the human rationales. All observations along the three inspected 'dimensions' hold for the OOD data, with attributions clearly being impacted (Figure 3), despite the lower influence in the AUC plausibility metric (Table 4).[3]

There is a disconnect between (over-)reliance on plausible features and OOD robustness. Despite the noticeable impact of constrained ER-C+ATTR and ER-C+IXG on attributions and their plausibility, Table 2 indicates a decline in classification performance. That is, the only ER strategies that resulted in models that rely more strongly on plausible tokens decrease OOD classification robustness. This is somewhat expected given what we observed before in Figure 5. However, it does lead us to conclude that the ER protocol should be re-considered, or at least, a bigger emphasis on trying to understand what 'amount' of increased plausibility is desirable, and where that information should be used while regularising the model, in order to better align the current expectations of improved OOD generalisation due to an increased reliance of ER models on plausible features.

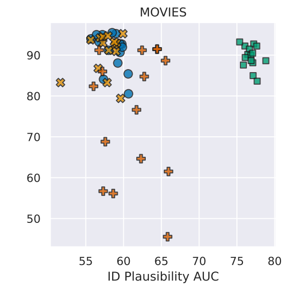
5.3 Predicting OOD performance

As shown in Figure 2, ER models exhibit a broad spread of OOD classification results. This highlights the importance of identifying the best-performing models, ideally without access to OOD data. Thus, we investigate whether it is possible to predict OOD classification performance from

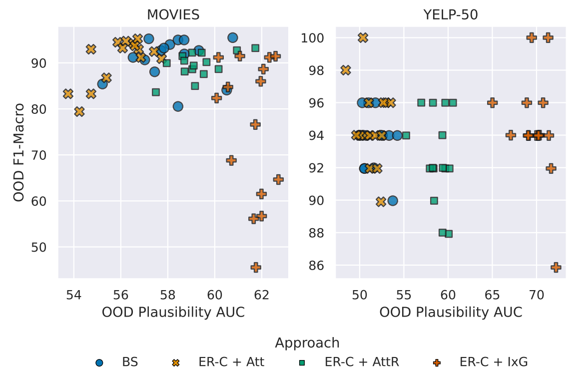
[3] We also find similar evidence for AUC per-layer across multiple attribution techniques (Appendix Figure 13).



(a) ID Classification



(b) ID Plaus. AUC w/ DX-C



(c) OOD Plausibility AUC w/ IxG

Figure 6: Relationship between ID/OOD predictors and OOD classification performance.

features potentially available for model selection.

In-domain classification and plausibility performance are not predictive of OOD improvements. Figures 6a and 6b show the association between in-domain (SST-Dev) predictors classification F1-Macro and plausibility AUC (computed with DX-C), and classification F1-Macro for Movies (OOD). In both cases there is no clear correlation between the in-domain metrics and OOD classification scores, with the result in Figure 6b further supporting the apparent disconnect between reliance on plausible features and OOD classification performance. We find a similar behavior for other approaches, OOD datasets, plausibility metrics and attribution techniques (Appendix F).

OOD plausibility performance is not predictive of OOD improvements. Given that we do not find in-domain classification performance or attribution plausibility to be predictive of OOD performance, we now investigate if there is any correlation at the OOD level. In order to test this, we use the Movies and Yelp-50 datasets. Figure 6c shows the association between OOD plausibility AUC (computed with IxG[4]) and classification F1-Macro, for both datasets. Similarly to what we observed in-domain, there is no correlation between OOD plausibility scores and its corresponding classification task performance.

The combination of both these results highlights the difficulty of predicting OOD results, and brings forth another challenge of the ER setup, where we need to make decisions about design choices based on in-domain performance, that does not seem to be linked to OOD generalisation.

6 Conclusion

In this work we take a step towards a better understanding of how explanation regularisation impacts a model beyond its classification performance. In particular, we aim to study both in- and out-of-domain settings, with a focus on how the plausibility of input attributions is affected, and what is the relationship between increased reliance on plausible tokens and robustness to OOD conditions.

We find that ER, unless constrained to meet a 'lower-bound' on the explanation loss computed with a global attribution technique, does not lead to models that effectively rely on plausible tokens. In fact, we observe a disconnect between reliance on plausible tokens and OOD robustness, with *(i)* OOD classification improvements not being predicted by increased plausibility; and *(ii)* OOD classification performance degrading when models are constrained to rely more on plausible tokens.

Our findings highlight relevant challenges for ER and suggest future research questions: *(i)* how to achieve a better balance between solving the classification task and constraining the model to attend to plausible tokens, *e.g.*, by learning which layers and attention heads of the Transformer model to regularise (see Fernandes et al. (2022)), or *(ii)* by allowing the model to learn to select which rationales to use during training (see Arous et al. (2021); Carton et al. (2022)); and *(iii)* how would input attributions for an ER model be impacted by using more informed, while still efficient, vector-based global guidance techniques as part of ER design choices (and not only as an analysis tool).

[4] MOVIES consists of long inputs, resulting in prohibitive runtime for batch sizes that do not result in OOM errors, on a Nvidia A100, when using DX and ALTI. Thus, we report IxG, as it can be computed for both datasets in the figure.

Limitations

In this section we identify some limitations of our work. However, we do note that we focus on the limitations that stem from design choices of our analysis, and not on the limitations we 'inherit' and that are part of the base problem, *e.g.*, whether plausibility should be a desired property of explanations (Jacovi and Goldberg, 2021).

Data. In terms of data, we identify two limitations. The first limitation related with data concerns the limited number of datasets per task that include annotations, making it difficult to analyse the impact of ER on the attributions on out-of-domain settings. Secondly, we use only the task of sentiment analysis to study the disconnect between ER's robustness to OOD conditions and the plausibility of attributions. This is also connected to the first discussed limitation — a full ER setup requires rationale-annotated data of an appreciable dimension for training, and also multiple OOD datasets, including at least one with rationale-annotated data. However, the questions we ask are deeply inherent to ER, and it seems unlikely that alternative tasks would be impacted differently.

Models. We use a single pre-trained model, BIGBIRD-ROBERTA-BASE (Zaheer et al., 2020). This choice follows existing work on ER (in particular applied to out-of-domain conditions), and allows evaluating datasets with long examples, but it does mean that we do not assess how ER impacts attributions for different pre-trained models, in particular, with varying number of parameters.

Attribution Techniques. Two of the techniques we use, ALTI (Ferrando et al., 2022) and DecompX (Modarressi et al., 2023), cannot be applied to long examples, leading either to prohibitive runtime, or to out-of-memory errors (on a single GPU NVIDIA A100), even with a batch size as low as 2. Thus, we cannot apply them to the OOD movies dataset, and have to limit our manually annotated split of Yelp to a moderate maximum sequence length. Yet, on Yelp, these represent a reasonable portion of the dataset's length distribution.

Ethical Considerations

This work studies the synergy between models that attribute more plausibly, *i.e.*, more aligned with humans, and improved OOD performance. We use existing datasets, that include human annotations of the relevant tokens that explain why an example is classified with a given label. These annotations might be biased, include mistakes, etc. By training a text classifier to become more aligned with those annotations we might further magnify the biases of the annotation process.

Acknowledgements

This research was done within the Mercury Machine Learning Lab, a collaboration between the University of Amsterdam, TU Delft, and Booking.com. Ivan Titov is supported by the Dutch National Science Foundation (NWO Vici VI.C.212.053). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197.

AI@Meta. 2024. Llama 3 model card.

Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. Marta: Leveraging human rationales for explainable text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 5868–5876.

Stephen P Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Samuel Carton, Surya Kanoria, and Chenhao Tan. 2022. What to learn, and how: Toward effective learning from rationales. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1075–1088.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307.

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2022. Unirex: A unified learning framework for language model rationale extraction. In *International Conference on Machine Learning*, pages 2867–2889. PMLR.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791.

Hila Chefer, Idan Schwartz, and Lior Wolf. 2022. Optimizing relevance maps of vision transformers improves robustness. *Advances in Neural Information Processing Systems*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. 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.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning.

Patrick Fernandes, Marcos Treviso, Danish Pruthi, André F.T. Martins, and Graham Neubig. 2022. Learning to scaffold: Optimizing model explanations for teaching. In *Advances in Neural Information Processing Systems*.

Javier Ferrando, Gerard I Gállego, and Marta R Costa-jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The eval4nlp shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2022. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159.

Reza Ghaeini, Xiaoli Z Fern, Hamed Shahbazi, and Prasad Tadepalli. 2019. Saliency learning: Teaching the model where to pay attention. In *Proceedings of NAACL-HLT*, pages 4016–4025.

Mareike Hartmann and Daniel Sonntag. 2022. A survey on improving nlp models with human explanations. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 40–47.

Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39, Dublin, Ireland. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

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

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.

Brihi Joshi, Aaron Chan, Ziyi Liu, Shaoliang Nie, Maziar Sanjabi, Hamed Firooz, and Xiang Ren. 2022. ER-test: Evaluating explanation regularization methods for language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3315–3336, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1611.07270*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2021. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

Mohammad Reza Ghasemi Madani and Pasquale Minervini. 2023. Refer: An end-to-end rationale extraction framework for explanation regularization. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 587–602.

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.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. Globenc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. 2020. Regularizing black-box models for improved interpretability. *Advances in Neural Information Processing Systems*, 33:10526–10536.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. 2022. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

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.

Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. 2022. Attcat: Explaining transformers via attentive class activation tokens. *Advances in neural information processing systems*, 35:5052–5064.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

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*, pages 1922–1933.

Lucas E. Resck, Marcos M. Raimundo, and Jorge Poco. 2024. Exploring the Trade-off Between Model Performance and Explanation Plausibility of Text Classifiers Using Human Rationales. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. 2020. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

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.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11349–11357.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5—rmsprop: Divide the gradient by a running average of its recent magnitude. Lecture 6.5. COURSERA: Neural networks for machine learning.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20.

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.

Zhuofan Ying, Peter Hase, and Mohit Bansal. 2022. Visfis: Visual feature importance supervision with right-for-the-right-reason objectives. In *Advances in Neural Information Processing Systems*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Attribution Techniques

We use a total of six attribution techniques in this work. They serve two main roles, acting as: *(i) guided attribution techniques*, when used as part of ER training to compute the explanation loss, $\mathcal{L}_{\text{expl}}$; and *(ii) non-guided attribution techniques*, when used to evaluate the impact of ER on input attributions. The attribution techniques we use are either *gradient-based* or *vector-based*. In the case of the latter, they differ on the components of the Transformer encoder architecture (Vaswani et al., 2017) that are used, and also on multiple design choices. Finally, all techniques provide *global* attributions, with the exception of attention, which is a *local* attribution technique.

Attention. Attention attributions correspond to the normalised top-layer attention weights. These are directly obtained from the Transformer model architecture. We use the values from the [CLS] token, averaged across heads.

Attention-Rollout. Attention-Rollout (Abnar and Zuidema, 2020) attributions are obtained by recursively aggregating attention scores. As in the original implementation, we incorporate the residual connection by adding the identity matrix ($\mathbf{I}$) to the original attention matrix ($\mathbf{A}^l$), followed by a normalization step, resulting in $\mathbf{A}^l_R = 0.5\mathbf{A}^l + 0.5\mathbf{I}$. We use the values from the [CLS] token, averaged across heads.

InputXGradient. INPUTXGRADIENT (Shrikumar et al., 2017) attributions are obtained by multiplying the input value by the corresponding gradient with respect to the output target label. We use the Captum package implementation (Kokhlikyan et al., 2020).

ALTI. ALTI (Ferrando et al., 2022) is a vector-based attribution technique based on the attention block decomposition introduced in Kobayashi et al. (2021). This decomposition factors in not only the attention weights, but also the value vectors, the output projection, as well as the first residual connection and layer normalization. The main difference lies in how the value of the contribution of a token at the layer level is computed — ALTI takes into account the output vector of the attention block and uses the L1 norm instead of the L2 norm. Furthermore, ALTI then aggregates the local attributions over the full model using a rollout-like approach, resulting in global attributions. We adapt the available code.[5]

DecompX. DecompX (Modarressi et al., 2023), similarly to ALTI, is a vector-based attribution technique based on the attention block decomposition by Kobayashi et al. (2021). The main differences are twofold: first, DecompX incorporates all the Transformer encoder components, including the feed-forward networks. Second, instead of using a recursive approach like rollout to aggregate the local attributions, DecompX propagates 'decomposed token representations' through the model. We also report DecompX-Classifier, which includes the classification head, and whose output is signed. We adapt the available code.[6]

B Impact of Local vs Global Guided Attributions in ER

We can use attention-rollout to show why we expect *global* attributions to be more impactful than the *local* counterparts when supervising a model with explanations. Rollout was introduced in Abnar and Zuidema (2020), and makes it possible to obtain global attributions by recursively aggregating vector-based local attributions, such as attention. By defining attention-rollout recursively, $\mathbf{a}_l = \rho(\mathbf{h}^l, \mathbf{h}^{l-1}; \mathbf{a}_{l-1})$, where $\mathbf{a}_l$ corresponds to attention-rollout weights at layer l, and $\mathbf{h}^l$ to the hidden states at layer l, we can write its gradient. Namely, for the lth layer, we get:

$$
\begin{aligned}
\frac{\partial \mathbf{a}_l}{\partial \theta} =& \frac{\partial}{\partial \theta}\rho(\mathbf{h}^l, \mathbf{h}^{l-1}; \mathbf{a}_{l-1}) \\
=& \frac{\partial}{\partial \mathbf{h}^l}\rho(\mathbf{h}^l, \mathbf{h}^{l-1}; \mathbf{a}_{l-1}) \times \frac{\partial \mathbf{h}^l}{\partial \theta} \\
&+ \frac{\partial}{\partial \mathbf{h}^{l-1}}\rho(\mathbf{h}^l, \mathbf{h}^{l-1}; \mathbf{a}_{l-1}) \times \frac{\partial \mathbf{h}^{l-1}}{\partial \theta} \\
&+ \frac{\partial}{\partial \mathbf{a}_{l-1}}\rho(\mathbf{h}^l, \mathbf{h}^{l-1}; \mathbf{a}_{l-1}) \times \underbrace{\frac{\partial \mathbf{a}_{l-1}}{\partial \theta}}_{\text{recursion}} .
\end{aligned}
\tag{3}
$$

By inspecting Equation 3, we can observe how the impact of guiding with a global vector-based technique goes beyond that of a local technique – there we have $\mathbf{a}_l = \rho(\mathbf{h}^l, \mathbf{h}^{l-1})$, meaning that the last recursion term of the gradient that propagates to the layers below would not be part of the computation. This difference seems to indicate that using a global attribution technique, such as

[5] https://github.com/mt-upc/transformer-contributions
[6] https://github.com/mohsenfayyaz/DecompX

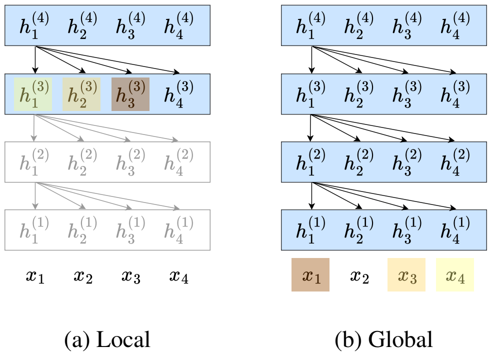


Figure 7: Illustration of *local* versus *global* attribution techniques as tools for ER. Local attributes to the input of the top-layer. Global attributes to the input tokens using the full model, and potentially constraining the model more strongly to follow human annotations.

attention-rollout, as E in the explanation loss will more strongly limit the model's ability to condition on features other than the rationale tokens. We show a visual interpretation of this difference in Figure 7.

C Attributions Techniques Faithfulness

To further validate our choice of attribution techniques, we assess the faithfulness of the used input attributions by computing normalised sufficiency and comprehensiveness (Carton et al., 2020). These are computed as follows:

$$\text{NullDiff}(x, \hat{y}) = \max(0, p(\hat{y}|x) - p(\hat{y}|x, 0)) \quad (4)$$

$$\text{NormSuff}(x, \hat{y}, \alpha) = \frac{\text{Suff}(x, \hat{y}, \alpha) - \text{Suff}(x, \hat{y}, 0)}{1 - \text{Suff}(x, \hat{y}, \alpha)} \quad (5)$$

$$\text{NormComp}(x, \hat{y}, \alpha) = \frac{\text{Comp}(x, \hat{y}, \alpha)}{\text{Comp}(x, \hat{y}, 1)}, \quad (6)$$

with $\hat{y} = \arg\max p(y|x)$. Both normalised sufficiency and comprehensiveness, computed with the top 1%, 20%, 40%, 60%, 80%, and 100% most attributed tokens, can be seen in Figure 8. As expected, both metrics improve the more top-attributed tokens, based on a given attribution technique, are used as input to the classifier. Furthermore, we find DecompX, DecompX-Classifier, and ALTI to perform the best. This corroborates the results reported on their respective works.

D Plausibility Metrics

We employ three metrics introduced in Fomicheva et al. (2021, 2022) to measure how well the input attributions match the human annotated tokens.

AUC Score. Computes the Area Under the Receiver Operating Characteristic Curve score, which considers multiple threshold values for the attributions with human annotations as the target label. This metric has been used in past works (DeYoung et al., 2020; Mathew et al., 2021; Fernandes et al., 2022; Resck et al., 2024).

Average Precision. As mentioned in Fomicheva et al. (2021), AUC scores might be optimistic when dealing with unbalanced data. That is relevant when working with rationales, where the annotated tokens might correspond to a small portion of the input. Thus, we report the average precision score, which summarizes the average-precision curve as a weighted average of the precision scores at the different thresholds,

$$\text{AP} = \sum_n (R_n - Rn - 1)\, P_n, \quad (7)$$

where P_n and R_n correspond to the precision and recall values at a given threshold.

Recall@k. Measures the average recall for a specific number of tokens k. It is calculated as

$$r@k = \frac{|\{x \in r_k(x) \;:\; x < k\}|}{N}, \quad (8)$$

where N is the number of annotated tokens for x, and $r(.)$ is a function that retrieves the rank of all annotated tokens.

E Input Attributions Correlations

As discussed in Section 5, one of the tools we use to study the impact of ER on the input model attributions is to measure changes in the correlation of attribution scores for a fixed technique across approaches, *e.g.*, how does attention correlate baseline vs baseline and baseline vs ER+ATT. Here, for each attribution technique, we iterate over all examples, sample two model versions, select the corresponding attributions for the BASELINE and approach we want to assess, and then compute the respective Kendall rank correlation coefficient[7]. We

[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kendalltau.html

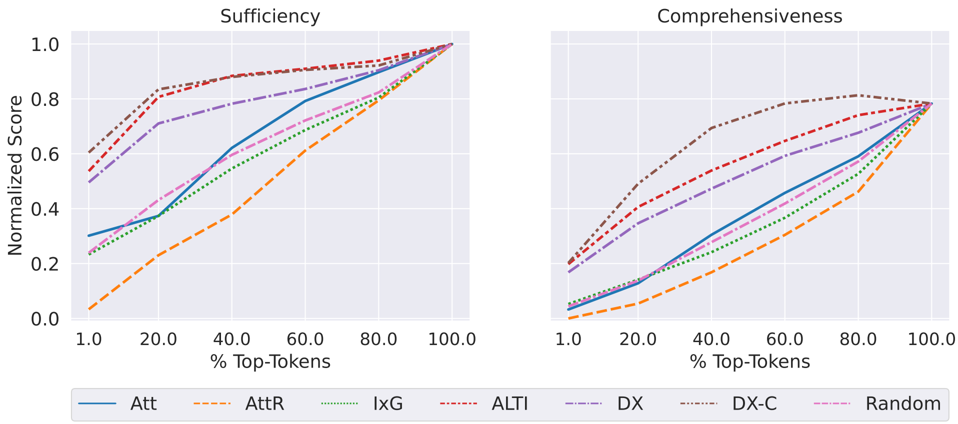


Figure 8: Normalised sufficiency and comprehensiveness (↑) scores for the top-x% tokens according to a given attribution technique, for SST-Dev, using the BASELINE approach.

show results for SST-Dev, Yelp-50, and Movies using all ER approaches and attribution techniques in Figure 9. The conclusions are as discussed in Section 5: ER techniques based on attention (ER+ATT and ER-C+ATT) impact only the attention attributions, while global constrained ER approaches (ER-C+ATTR and ER-C+IXG) are able to impact all reported attribution techniques.

Alternatively, it is also possible to compare correlation coefficients for a pair of attribution techniques for a given choice of training objective (*e.g.*, the correlation between attention and attention-rollout for baseline or ER+ATT) and how those correlations change as we vary the training objective. We compute average Kendall rank correlation values and report them in Figure 10. The obtained results agree with our previous findings: *(i)* attention as a guided attribution technique (ER+ATT and ER-C+ATT) impacts mostly only attention attributions; *(ii)* global guided attribution techniques have little impact when used in joint ER (ER+ATTR and ER+IXG); *(iii)* only the constrained ER approaches with global guided attributions techniques are able to clearly impact most attributions; and *(iv)* for OOD most patterns are the same, despite less noticeable. Note how this correlation analysis is different from the previous. First, here we inspect averages (the number of data points makes it unfeasible to report histograms of correlation coefficients), which has the potential to be misleading. Second, we correlate across attribution techniques instead of correlating the same attribution across approaches, which leads to observations that are more complex and difficult to interpret.

F Predicting OOD Performance

To complement the results discussed in Section 5.3, we present findings on a broader range of OOD datasets and plausibility metrics criteria, also including joint ER+ATT and ER+ATTR results. Results are shown in Figure 14. Similarly to Section 5.3, no in-domain criteria exhibits correlation with OOD classification performance.

G Data

For in-domain and training data we use SST-2 (Socher et al., 2013), following the heuristic algorithm proposed in Carton et al. (2020) to obtain instance-level rationales[8]. For out-of-domain we use: Amazon-Reviews[9] (Hou et al., 2024), IMBD[10] (Maas et al., 2011), Movies[11] (Zaidan and Eisner, 2008; DeYoung et al., 2020), and Yelp[12] (Zhang et al., 2015). From these OOD datasets only Movies includes human annotated rationales. Some OOD evaluation datasets have several thousand examples. In those cases we sample 5,000 examples. More details on the final data are described in Table 6. All data is in English.

[8] https://github.com/BoulderDS/evaluating-human-rationales/blob/master/scripts/download_and_process_sst.py

[9] https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023

[10] https://huggingface.co/datasets/stanfordnlp/imdb

[11] https://huggingface.co/datasets/eraser-benchmark/movie_rationales

[12] https://huggingface.co/datasets/fancyzhx/yelp_polarity

	Train	Dev	Test	% Positive	Avg. / Max Length	Source	Rationales
SST (Socher et al., 2013)	6,920	872	1,821	≈ 50 %	≈ 25 / ≈ 65	Movies	✓
MOVIES (Zaidan and Eisner, 2008; DeYoung et al., 2020)	-	-	399*	50.0 %	786 / 1024	Movies	✓
IMDB (Maas et al., 2011)	-	-	5,000	50.7 %	288 / 1024	Movies	×
AMAZON (MOVIES AND TV) (Hou et al., 2024)	-	-	5,000	77.5 %	107 / 1024	Products	×
YELP (Zhang et al., 2015)	-	-	5,000	50.6 %	168 / 1024	Businesses	×
YELP-50**	-	-	50	50.0 %	84 / 151	Businesses	✓

Table 6: Sentiment analysis datasets, including percentage of examples labeled as positive, average and maximum length of inputs (after truncating), source of the data, and whether rationales are provided. * – This includes both dev and test examples. ** – This corresponds to the subset of annotated data described in Section G.1.

	Average F1-Macro
BASELINE	93.71 ± 1.62
ER+ATT	93.98 ± 1.80
ER+ATTR	94.24 ± 2.07
ER+IXG	93.30 ± 1.91
ER-C+ATT	94.64 ± 2.41
ER-C+ATTR	92.64 ± 2.61
ER-C+IXG	94.51 ± 3.17

Table 7: F1-Macro scores (↑) for YELP-50.

G.1 Extra OOD Annotated Data

We conduct our OOD analysis on the Movies dataset, an available sentiment analysis dataset with human rationale annotations. To further support our claims with respect to increased robustness to OOD conditions, we annotated a split of 50 examples of the Yelp dataset. We first filter out examples that will result in more than 150 subtokens when using the tokenizer from the pre-trained encoder model we use during training (so that we can compute attributions with all available metrics, in particular ALTI (Ferrando et al., 2022) and DecompX (Modarressi et al., 2023), which we find to result in prohibitive runtime or OOM errors with long sequences). Then, we sample 55 examples at random from those, tokenize them using the English tokenizer from spaCy (Honnibal et al., 2020), and manually annotate the first five, selecting all sequences of text that offer evidence for the gold label. These are used as few-shot examples for a prompt to META-LLAMA-3-8B-INSTRUCT[13] (AI@Meta, 2024) that outputs automatic annotations for the remaining 50 examples. Finally, we post-edit the automatic annotations whenever necessary.

Average model classification performance for this split of data is shown in Table 7.

[13] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

H Hyperparameters and Training

Optimizer and Scheduler. We use AdamW (Loshchilov and Hutter, 2018) as our optimizer, with β values set to (0.9, 0.98) and the weight decay coefficient set to 0. For the learning rate we use a linear scheduler, with 10% of training steps as warm-up. For constrained optimisation parameters we use RMSprop (Tieleman and Hinton, 2012).

Hyperparameter Selection. We choose the combination of learning rate and maximum number of training epochs (since we use a linear scheduler with 10% of training steps as warm-up) with the lowest cross-entropy loss over three runs. For the explanation regularised approach we explore the same hyperparameters, plus the λ weight. We select based on cross-entropy loss so that different values of λ can be directly compared.

For the constrained approach, we first train a model (3 seeds) using $\mathcal{L} = \mathcal{L}_{\text{expl}}$ and use the average minimum explanation train and validation explanation losses to guide our choice of bounds b_{train} and b_{val}. For b_{val} we use the average of the minimum validation loss. For b_{train} we use a value close to 1.5 times the average of the minimum training loss. Then, we choose the combination of learning rate and constrained optimizer learning rate that minimizes the average cross entropy loss.

For the baseline we explore learning rate $\in \{2, 3, 5 \times 10^{-5}\}$ and maximum number of epochs $\in \{15, 25\}$. For the joint approach we explore the same space as the baseline model and $\lambda \in \{0.6, 1.0, 1.4\}$, following a set of choices aligned with previous work on ER for OOD robustness. For the constrained approach we explore learning rate $\in \{2, 3, 5 \times 10^{-5}\}$ and constrained learning rate $\in \{1 \times 10^{-1}, 5 \times 10^{-2}\}$. All experiments use a batch size of 32. Final choices can be seen in Table 8. The experiments with multiple lambdas uses the same choices as the baseline model.

Baseline	Value
Learning Rate	3×10^{-5}
Train Epochs	25
Joint Attention	**Value**
Learning Rate	2×10^{-5}
Train Epochs	25
λ	1.0
Joint Rollout	**Value**
Learning Rate	3×10^{-5}
Train Epochs	15
λ	1.0
Joint IxG	**Value**
Learning Rate	2×10^{-5}
Train Epochs	25
λ	1.0
Constrained Attention	**Value**
Learning Rate	2×10^{-5}
Train Epochs	25
Constrained Learning Rate	5×10^{-2}
b_{train}	0.035
b_{val}	0.031
Constrained Rollout	**Value**
Learning Rate	3×10^{-5}
Train Epochs	25
Constrained Learning Rate	1×10^{-1}
b_{train}	0.030
b_{val}	0.031
Constrained IxG	**Value**
Learning Rate	2×10^{-5}
Train Epochs	25
Constrained Learning Rate	1×10^{-1}
b_{train}	0.35
b_{val}	0.35

Table 8: Hyperparameters choices for all approaches.

Training. During training we choose a model checkpoint based on average validation loss for the baseline and the *joint* explanation regularisation approach. This follows Joshi et al. (2022) and ensures that the explanation loss of the ER approach directly influences model selection.

For constrained optimization, we choose the model checkpoint with the lowest validation cross-entropy loss, provided that the validation explanation loss is $\mathcal{L}_{\text{expl}_{\text{val}}} < 1.1 \times b_{\text{val}}$. This ensures that we compare model checkpoints where the guided attribution technique is learning to predict the annotated rationales, according to the defined bound. Selecting a checkpoint based on the same criteria would have not been possible for the joint approach, as none of the runs converges to an explanation loss value that meets the defined validation bound.

Unless mentioned otherwise, we report results over 15 seeds.[14] All our experiments are developed using a single Nvidia A100 40GB GPU, and implemented with PyTorch (Paszke et al., 2019) and PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019).

Attribution Techniques. Following Joshi et al. (2022), we scale the attributions $E(C_\theta, x)$ by 100 when computing the explanation loss and renormalize them with softmax. We take it as part of the approaches that use attention as the guided attribution technique, ER+ATT and ER-C+ATT. For INPUTXGRADIENT we keep all default choices of the Captum package. The output attributions are aggregated into token-level attributions via sum, and normalised with the L2 norm. For DECOMPX/DECOMPX-C[15], and ALTI[16] we keep all default choices part of the original work. For attributions techniques that output 'signed' attribution scores, *i.e.*, IXG and DX-C, we take the absolute value of the attribution score.

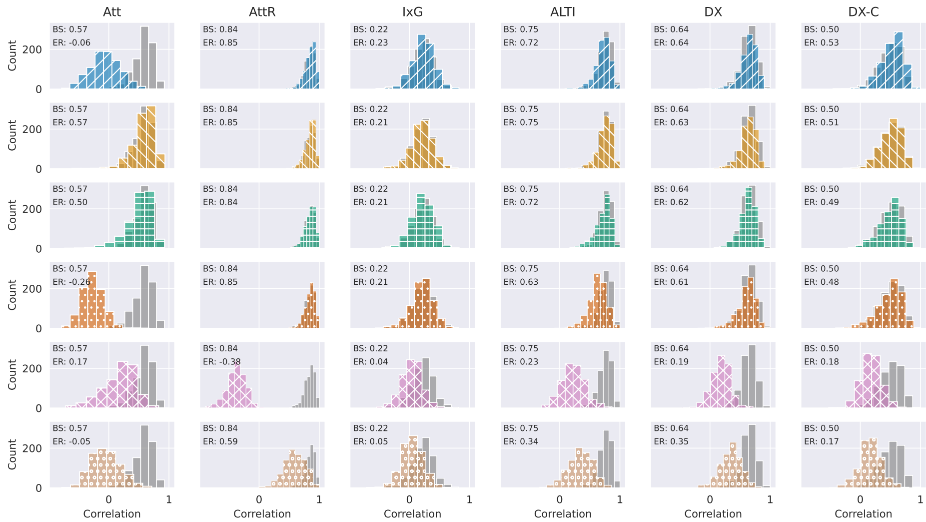
Classifier. Following Joshi et al. (2022), we use a pre-trained Transformer encoder model, GOOGLE/BIGBIRD-ROBERTA-BASE[17] (Zaheer et al., 2020), followed by a linear layer that uses as input the top-layer representation of the [CLS] token, with TANH as the non-linearity, and no classifier dropout.

[14] We found ER-C+IXG unstable to train, with only around one in three seeds minimising the CE loss while also minimising the explanation loss. Thus, it required training more models. In practice, we still report results over 15 seeds.

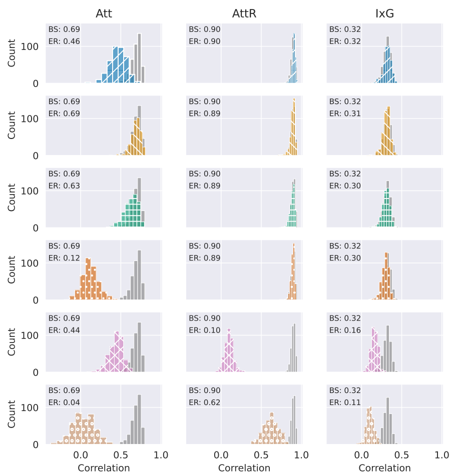
[15] https://github.com/mohsenfayyaz/DecompX

[16] https://github.com/mt-upc/transformer-contributions

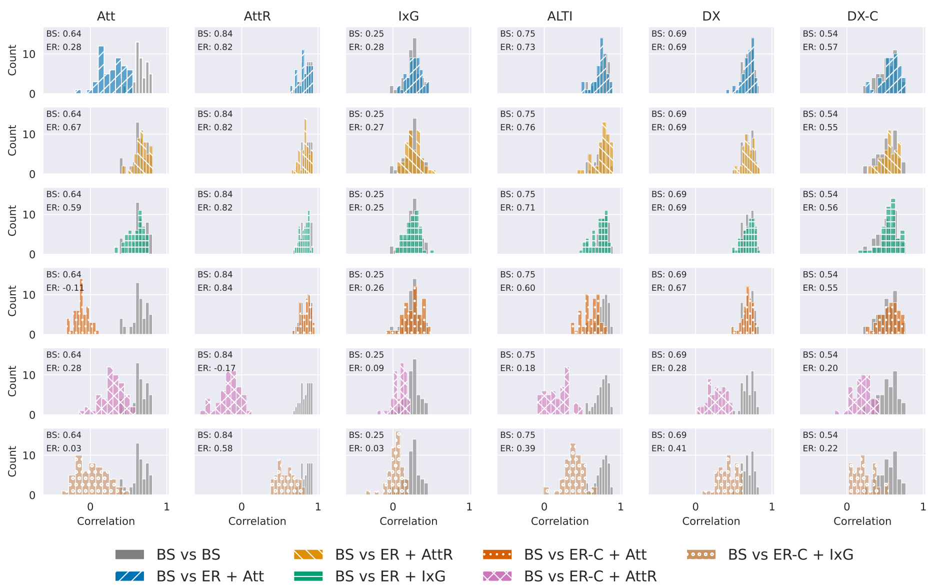
[17] https://huggingface.co/google/bigbird-roberta-base



(a) SST-Dev (ID)



(b) Movies (OOD)



(c) Yelp-50 (OOD)

Figure 9: Average Kendall Rank correlation between attribution techniques of the different approaches. The BS (Baseline) and ER (corresponding ER approach) text corresponds to the average correlation values.

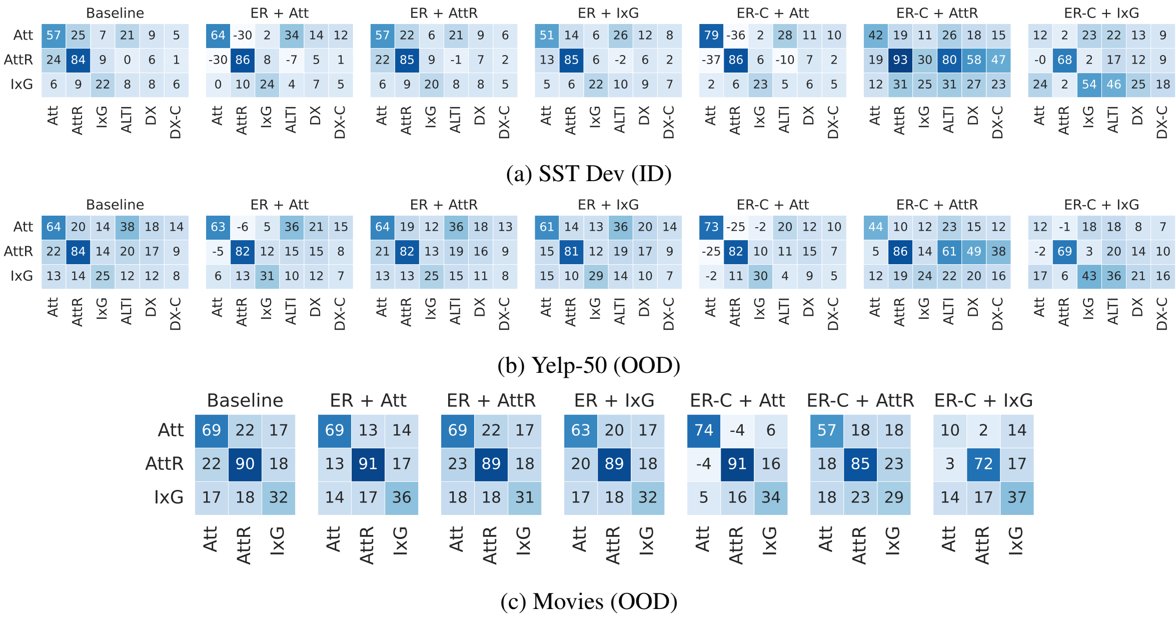


Figure 10: Average Kendall Rank correlation between attribution techniques for the different approaches.

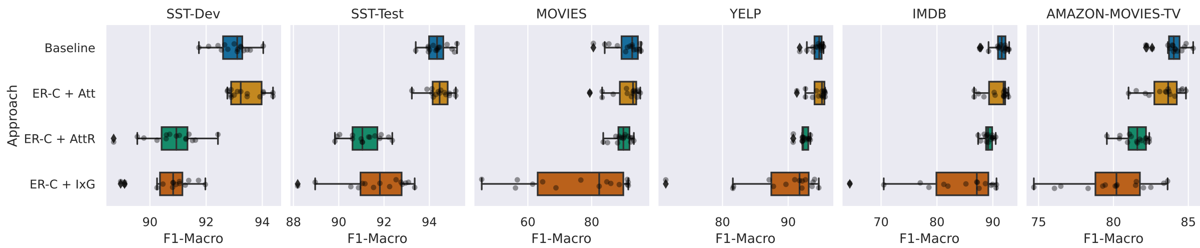


Figure 11: F1-Macro scores (↑). ER-C + ATT uses attention as the guided attribution technique, ER-C + ATTR uses attention-rollout, and ER-C + IXG uses INPUTXGRADIENT. Results correspond to 15 seeds.

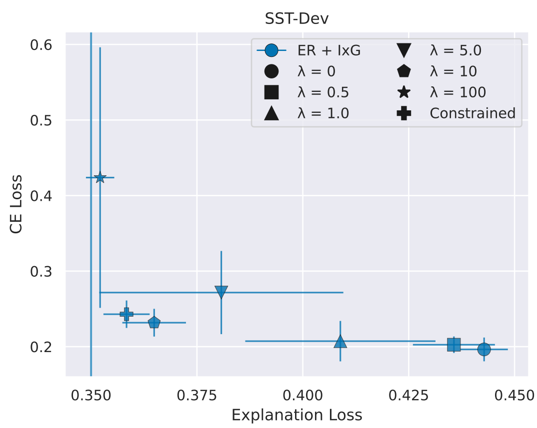
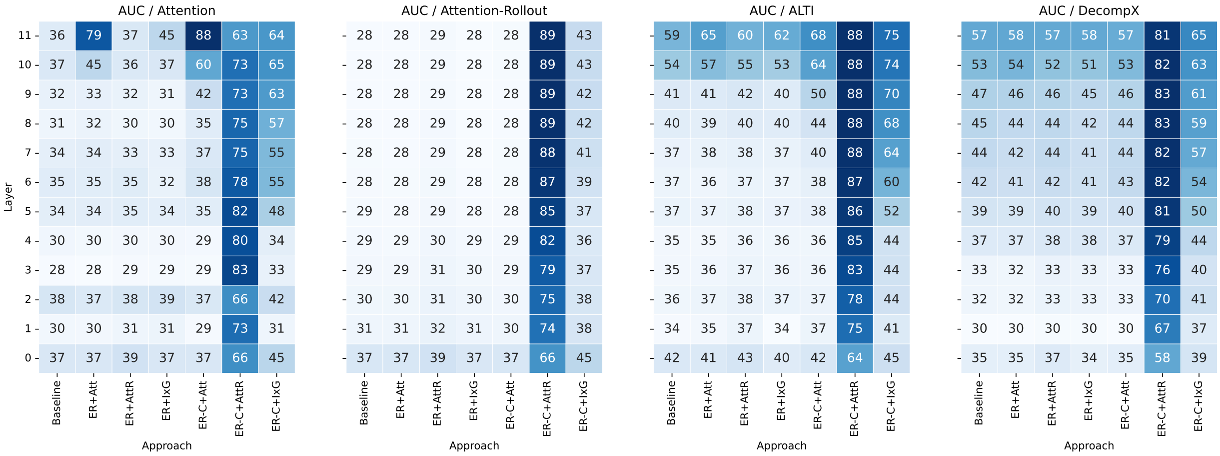
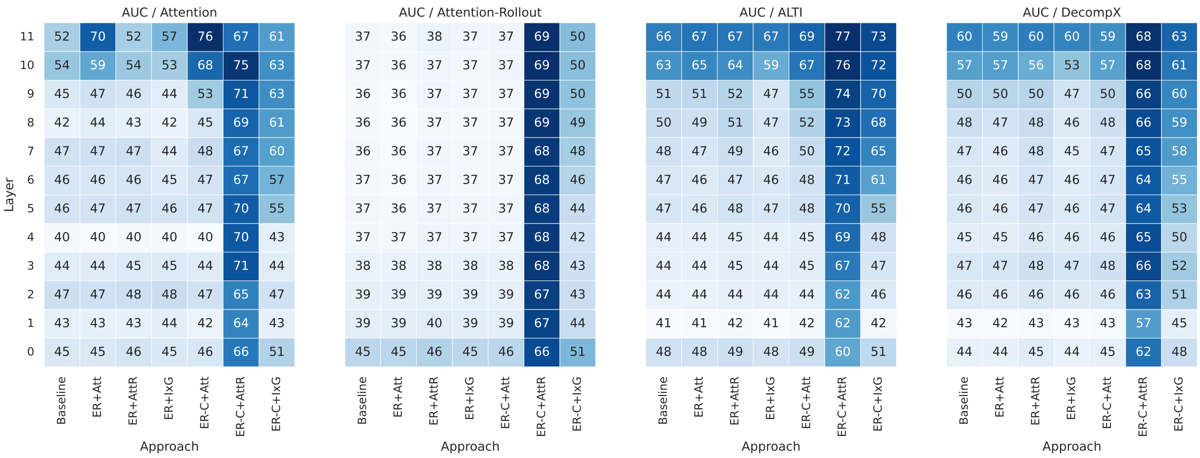


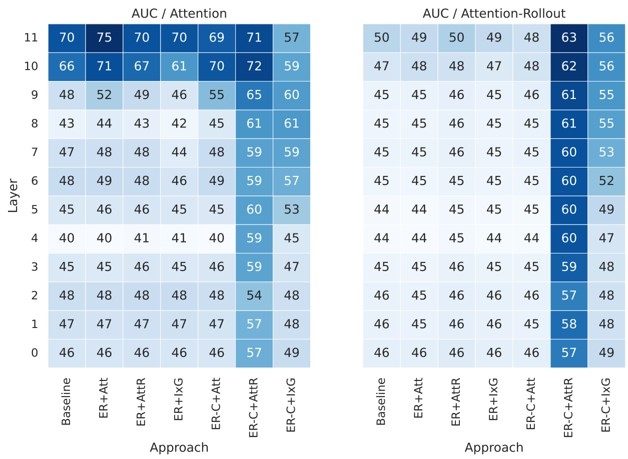
Figure 12: Cross-Entropy vs Explanation Loss for SST-Dev, using ER+IXG. The vertical line shows the validation explanation-loss bound. Each point corresponds to the average of 5 runs, and the error bars to the standard deviation for each loss. λ ranges from 0 (No-ER) to 100.



(a) SST-Dev (ID)

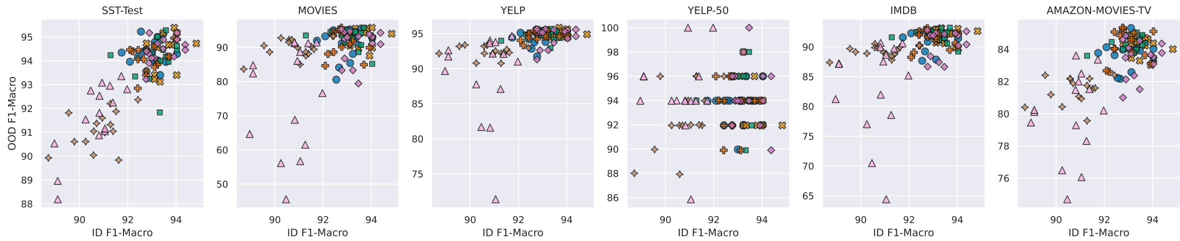


(b) Yelp-50 (OOD)

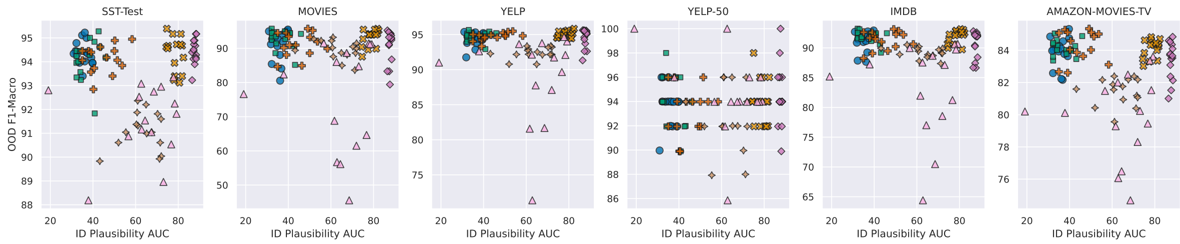


(c) Movies (OOD)

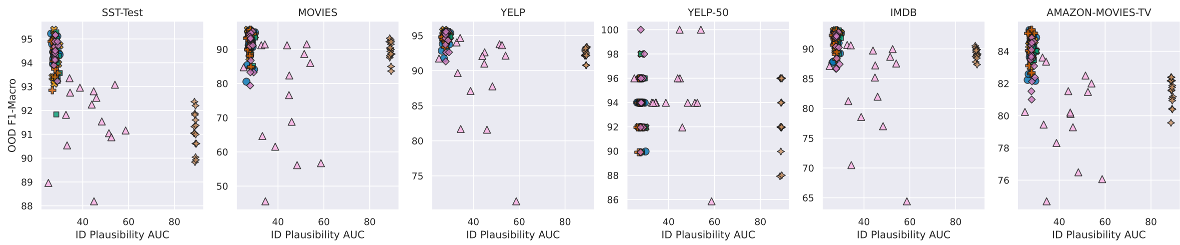
Figure 13: Average AUC plausibility scores per-layer (↑). We report only techniques that output per-layer attributions.



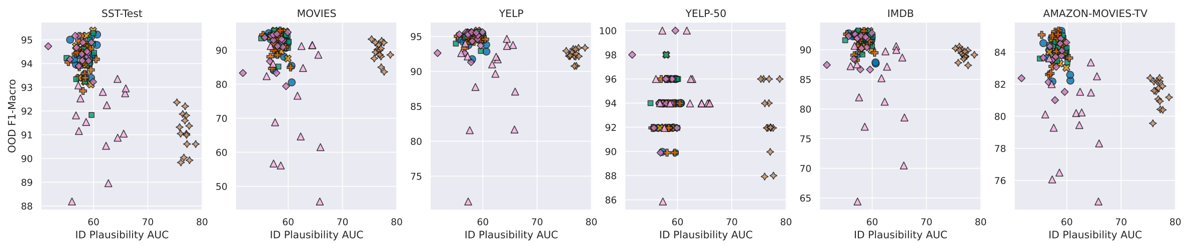
(a) SST-Dev Classification as Predictor



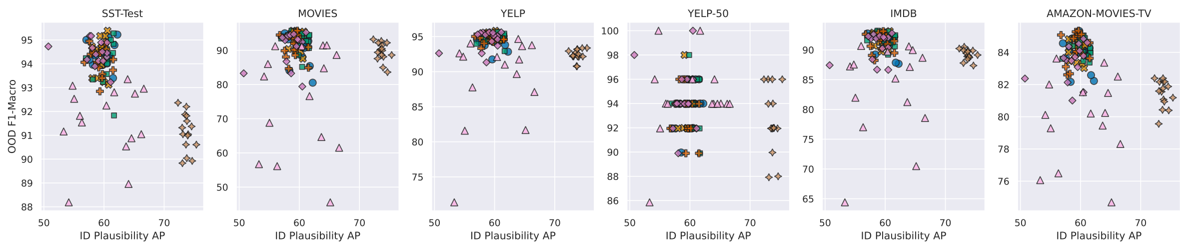
(b) SST-Dev Plausibility as Predictor - AUC with Attention



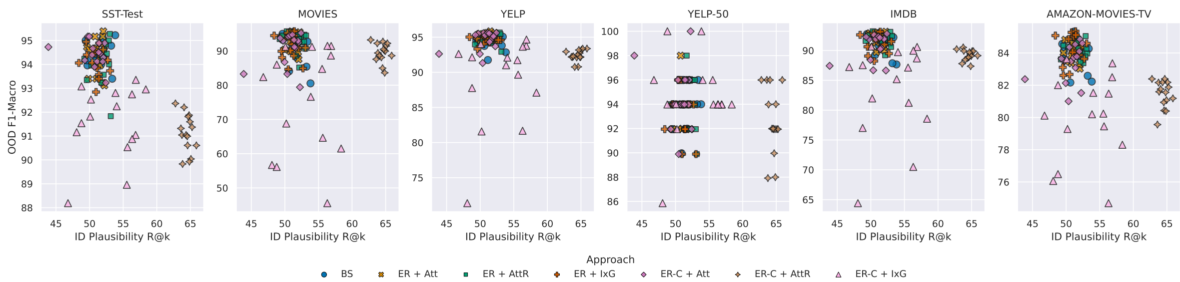
(c) SST-Dev Plausibility as Predictor - AUC with Attention-Rollout



(d) SST-Dev Plausibility as Predictor - AUC with DecompX-Classifier



(e) SST-Dev Plausibility as Predictor - Average Precision with DecompX-Classifier



(f) SST-Dev Plausibility as Predictor - Recall@k with DecompX-Classifier

Figure 14: Scatter plot showing the relationship between in-domain (SST-Dev) measurements (F1-Macro classification scores and plausibility scores computed with different metrics and attribution techniques) for multiple datasets.

	AUC						AP						R@k					
	Att	AttR	IxG	ALTI	DX-C	DX	Att	AttR	IxG	ALTI	DX-C	DX	Att	AttR	IxG	ALTI	DX-C	DX
SST-Dev (ID)																		
Baseline	35.6 ± 2.9	27.9 ± 1.1	50.7 ± 1.5	59.0 ± 2.8	58.5 ± 1.6	57.5 ± 1.7	39.8 ± 2.0	35.6 ± 0.4	50.7 ± 1.0	61.2 ± 2.2	59.8 ± 1.6	58.3 ± 1.7	32.0 ± 1.9	28.0 ± 0.6	42.7 ± 0.9	52.3 ± 2.2	51.4 ± 1.3	50.4 ± 1.4
ER + Att	78.6 ± 2.8	27.9 ± 0.8	50.6 ± 1.8	64.5 ± 2.4	58.4 ± 0.9	57.8 ± 1.6	76.7 ± 2.2	35.7 ± 0.4	50.5 ± 1.5	64.1 ± 1.3	59.5 ± 0.9	58.0 ± 1.4	67.9 ± 2.3	28.3 ± 0.5	42.5 ± 1.2	55.5 ± 1.4	51.2 ± 0.9	50.3 ± 1.3
ER + AttR	37.0 ± 4.7	28.7 ± 0.6	50.8 ± 2.1	60.2 ± 3.0	58.6 ± 1.4	57.4 ± 2.0	40.8 ± 3.1	35.9 ± 0.3	50.5 ± 1.9	61.9 ± 2.1	60.2 ± 1.4	58.6 ± 2.0	32.9 ± 3.6	28.4 ± 0.4	42.8 ± 1.4	53.3 ± 2.1	51.8 ± 1.2	50.8 ± 1.6
ER + IxG	44.6 ± 7.2	27.6 ± 0.7	53.8 ± 4.7	61.7 ± 4.6	58.0 ± 1.2	57.6 ± 1.7	47.0 ± 5.5	35.5 ± 0.2	53.7 ± 4.3	62.6 ± 3.4	59.3 ± 1.3	57.8 ± 1.5	39.3 ± 5.3	27.9 ± 0.4	45.7 ± 4.0	53.8 ± 3.4	51.0 ± 1.2	50.0 ± 1.4
ER-C + Att	87.6 ± 0.7	28.2 ± 0.5	51.5 ± 1.4	68.3 ± 3.6	57.6 ± 2.0	56.8 ± 1.5	84.8 ± 0.5	35.7 ± 0.2	51.2 ± 1.2	65.8 ± 2.1	58.6 ± 2.5	56.6 ± 1.6	76.3 ± 0.6	28.3 ± 0.4	43.2 ± 1.0	57.9 ± 2.2	50.6 ± 2.1	49.2 ± 1.4
ER-C + AttR	63.2 ± 8.4	89.1 ± 0.3	67.6 ± 1.5	87.9 ± 0.6	76.9 ± 0.9	81.3 ± 0.6	63.9 ± 7.7	85.6 ± 0.5	61.8 ± 1.1	84.9 ± 0.6	73.7 ± 0.7	77.0 ± 0.5	55.1 ± 7.6	77.1 ± 0.6	53.8 ± 1.1	76.5 ± 0.7	64.5 ± 0.8	67.9 ± 0.6
ER-C + IxG	63.9 ± 16	42.9 ± 9.5	81.5 ± 2.8	74.8 ± 3.1	61.3 ± 3.6	64.8 ± 4.1	64.2 ± 14	42.4 ± 4.9	83.2 ± 2.5	75.3 ± 3.0	60.4 ± 4.9	62.1 ± 5.6	55.5 ± 14	37.1 ± 6.4	74.0 ± 2.7	67.3 ± 2.7	53.1 ± 3.9	55.7 ± 4.6
$\mathcal{L}_{\text{expl}}$ (A)	89.0 ± 0.4	29.6 ± 0.5	53.2 ± 1.6	78.2 ± 1.2	52.0 ± 0.7	53.0 ± 0.8	87.4 ± 0.3	36.0 ± 0.1	51.8 ± 0.9	73.8 ± 0.2	50.9 ± 1.0	50.0 ± 1.6	79.2 ± 0.3	28.6 ± 0.0	43.6 ± 0.6	65.0 ± 0.2	44.0 ± 0.5	43.6 ± 1.8
$\mathcal{L}_{\text{expl}}$ (R)	81.1 ± 2.4	89.4 ± 0.2	72.9 ± 1.4	88.2 ± 0.5	77.1 ± 2.5	83.3 ± 1.6	76.3 ± 1.8	86.7 ± 0.4	65.4 ± 0.6	84.0 ± 0.8	72.3 ± 1.6	77.6 ± 1.4	69.6 ± 2.1	78.6 ± 0.4	57.5 ± 0.6	76.3 ± 0.9	63.5 ± 1.6	69.4 ± 1.7
$\mathcal{L}_{\text{expl}}$ (IxG)	69.2 ± 6.3	57.4 ± 13	84.3 ± 0.5	81.0 ± 1.0	65.2 ± 1.3	66.1 ± 9.1	67.9 ± 12	51.0 ± 9.4	85.4 ± 0.6	82.4 ± 0.9	62.9 ± 2.6	66.1 ± 11	59.1 ± 12	48.4 ± 12	76.9 ± 1.0	73.5 ± 1.1	55.8 ± 3.3	59.3 ± 8.8
Movies (OOD)																		
Baseline	70.0 ± 2.5	49.6 ± 1.2	58.1 ± 1.4	-	-	-	27.7 ± 2.2	18.5 ± 0.3	22.6 ± 0.8	-	-	-	28.2 ± 2.7	16.6 ± 0.4	22.8 ± 1.0	-	-	-
ER + Att	74.6 ± 2.0	49.1 ± 0.7	58.1 ± 0.7	-	-	-	34.9 ± 2.9	18.4 ± 0.2	22.7 ± 0.5	-	-	-	35.3 ± 2.9	16.5 ± 0.2	23.1 ± 0.6	-	-	-
ER + AttR	70.1 ± 1.7	50.2 ± 1.1	58.0 ± 1.0	-	-	-	27.5 ± 1.3	18.7 ± 0.3	22.5 ± 0.5	-	-	-	28.0 ± 1.5	16.8 ± 0.4	22.7 ± 0.6	-	-	-
ER + IxG	69.9 ± 2.8	49.0 ± 1.3	59.5 ± 1.5	-	-	-	29.9 ± 2.6	18.4 ± 0.3	23.6 ± 1.0	-	-	-	30.7 ± 2.8	16.5 ± 0.4	24.1 ± 1.3			
ER-C + Att	68.8 ± 1.6	48.1 ± 0.6	56.0 ± 1.2	-	-	-	31.7 ± 1.6	18.2 ± 0.1	21.5 ± 0.6	-	-	-	32.6 ± 1.4	16.3 ± 0.2	21.6 ± 0.8	-	-	-
ER-C + AttR	71.1 ± 2.0	62.8 ± 1.3	59.3 ± 1.1	-	-	-	31.3 ± 2.4	26.1 ± 0.8	23.1 ± 0.7	-	-	-	32.1 ± 2.7	26.6 ± 1.0	23.6 ± 0.9	-	-	-
ER-C + IxG	56.7 ± 11	56.1 ± 3.6	61.6 ± 0.8	-	-	-	23.6 ± 5.6	20.6 ± 1.4	26.6 ± 0.8	-	-	-	22.7 ± 7.3	19.8 ± 2.1	28.2 ± 0.8			
$\mathcal{L}_{\text{expl}}$ (A)	61.0 ± 2.4	48.2 ± 1.5	48.0 ± 0.3	-	-	-	25.6 ± 0.6	18.0 ± 0.3	18.3 ± 0.1	-	-	-	27.4 ± 0.3	16.1 ± 0.4	17.0 ± 0.0	-	-	-
$\mathcal{L}_{\text{expl}}$ (R)	59.4 ± 1.3	62.5 ± 0.4	55.2 ± 0.8	-	-	-	23.8 ± 0.3	25.8 ± 0.2	21.7 ± 0.4	-	-	-	22.7 ± 0.8	26.6 ± 0.2	21.2 ± 0.5	-	-	-
$\mathcal{L}_{\text{expl}}$ (IxG)	51.8 ± 7.3	56.9 ± 4.0	59.8 ± 1.0	-	-	-	20.9 ± 2.6	21.2 ± 2.3	25.1 ± 0.6	-	-	-	19.6 ± 4.8	20.9 ± 3.6	26.8 ± 0.7			
Yelp-50 (OOD)																		
Baseline	51.9 ± 2.5	37.1 ± 1.1	51.5 ± 1.4	66.2 ± 2.1	59.5 ± 2.0	60.3 ± 1.4	28.5 ± 2.2	21.8 ± 0.3	31.1 ± 1.2	48.4 ± 3.9	41.7 ± 2.5	39.5 ± 2.1	23.2 ± 3.0	14.3 ± 0.6	26.3 ± 1.4	41.7 ± 2.7	36.6 ± 2.4	36.3 ± 1.7
ER + Att	69.7 ± 2.0	36.4 ± 1.1	50.7 ± 1.2	67.4 ± 1.9	59.1 ± 0.9	59.3 ± 1.5	53.1 ± 1.8	21.7 ± 0.3	30.8 ± 1.2	50.5 ± 2.2	40.6 ± 1.5	38.0 ± 1.7	46.8 ± 2.2	14.0 ± 0.5	25.6 ± 1.4	43.9 ± 2.1	35.7 ± 1.4	35.2 ± 1.7
ER + AttR	52.5 ± 3.0	37.6 ± 1.2	51.7 ± 0.8	67.1 ± 1.5	59.8 ± 1.2	60.5 ± 1.3	28.9 ± 2.7	22.0 ± 0.3	31.4 ± 1.2	49.8 ± 2.8	42.6 ± 1.7	40.5 ± 2.1	23.5 ± 3.9	14.4 ± 0.7	26.6 ± 1.3	42.8 ± 2.1	37.0 ± 1.3	36.7 ± 1.7
ER + IxG	56.5 ± 3.3	37.3 ± 1.0	53.6 ± 2.4	66.6 ± 2.6	59.2 ± 1.5	59.9 ± 1.7	34.9 ± 4.5	21.9 ± 0.3	33.4 ± 2.8	50.1 ± 4.1	41.5 ± 2.4	39.3 ± 2.4	30.0 ± 5.1	14.3 ± 0.7	28.9 ± 3.1	42.9 ± 3.8	36.2 ± 2.1	35.7 ± 2.0
ER-C + Att	76.0 ± 1.2	37.4 ± 1.2	51.3 ± 1.4	69.4 ± 2.7	58.6 ± 2.9	58.9 ± 2.5	56.4 ± 1.5	21.9 ± 0.3	31.1 ± 1.2	49.5 ± 2.7	39.2 ± 3.4	36.8 ± 2.8	50.9 ± 1.6	14.4 ± 0.7	26.2 ± 1.6	43.6 ± 2.8	35.4 ± 3.2	34.5 ± 2.9
ER-C + AttR	67.2 ± 5.8	69.4 ± 1.5	58.8 ± 1.4	76.7 ± 1.6	65.5 ± 1.5	67.9 ± 1.6	49.4 ± 7.5	49.5 ± 1.5	37.8 ± 1.6	58.3 ± 1.8	47.7 ± 1.3	49.0 ± 1.3	43.4 ± 7.1	43.8 ± 1.4	34.1 ± 1.6	51.8 ± 1.8	42.3 ± 1.5	43.4 ± 1.5
ER-C + IxG	60.8 ± 13	50.5 ± 6.3	69.8 ± 1.8	73.2 ± 2.6	59.2 ± 3.5	63.0 ± 5.1	41.3 ± 13	27.5 ± 3.3	52.3 ± 0.9	54.4 ± 2.8	39.4 ± 4.4	41.1 ± 5.7	35.3 ± 13	23.5 ± 5.5	47.4 ± 1.3	48.8 ± 2.8	35.7 ± 4.0	39.0 ± 5.5
$\mathcal{L}_{\text{expl}}$ (A)	70.1 ± 2.1	38.5 ± 1.9	47.1 ± 0.7	62.5 ± 1.3	51.3 ± 1.3	49.4 ± 3.1	50.3 ± 2.5	21.8 ± 0.7	27.9 ± 0.4	41.2 ± 1.1	30.0 ± 1.0	27.8 ± 1.6	45.4 ± 1.7	14.6 ± 1.1	23.6 ± 0.4	36.3 ± 1.3	25.9 ± 1.0	23.1 ± 2.3
$\mathcal{L}_{\text{expl}}$ (R)	62.6 ± 2.9	68.6 ± 1.0	56.8 ± 1.7	66.5 ± 1.3	59.0 ± 1.7	61.6 ± 0.2	39.3 ± 2.9	47.1 ± 1.8	35.2 ± 1.4	43.8 ± 1.7	37.1 ± 1.3	39.0 ± 0.4	35.7 ± 4.0	42.1 ± 2.0	31.9 ± 1.2	39.5 ± 2.1	33.7 ± 1.4	35.6 ± 0.8
$\mathcal{L}_{\text{expl}}$ (IxG)	55.2 ± 12	58.3 ± 8.3	68.6 ± 0.7	72.2 ± 1.7	58.8 ± 1.6	60.7 ± 6.0	37.7 ± 10	33.4 ± 7.5	49.5 ± 0.2	53.6 ± 2.2	38.2 ± 2.0	41.1 ± 8.1	31.6 ± 11	32.2 ± 10	44.9 ± 1.0	47.1 ± 2.1	34.9 ± 2.5	38.3 ± 7.4

Table 9: **Attributions to Annotated Tokens**. AUC, Average Precision and Recall@k Scores (↑) for SST-Dev (ID), Movies (OOD), and Yelp-50 (OOD) datasets, for the Baseline (Baseline), ER model guided with Attention (ER + Att), ER model guided with Attention-Rollout (ER + AttR), ER model guided with InputXGradient (ER + IxG) and the corresponding constrained counterparts (ER-C + X). Attributions are computed using Attention (Att), Attention + Rollout (AttR), InputXGradient (IxG), ALTI Aggregated (ALTI), and DecompX (with) Classifier (DX and DX-C) as the attribution techniques. For Movies, we report only the techniques that can be applied to longer sequences.