BERT, are you paying attention? Attention regularisation with human-annotated rationales

Elize Herrewijnen[1,3], Dong Nguyen[1], Floris Bex[1,2,3], Albert Gatt[1]
[1]Department of Information & Computing Sciences, Utrecht University, Utrecht, Netherlands
[2]School of Law, Utrecht University Utrecht, Netherlands
[3]National Police Lab AI, Netherlands Police, Driebergen, Netherlands
Correspondence: e.herrewijnen@uu.nl

Abstract

Attention regularisation aims to supervise the attention patterns in language models like BERT. Various studies have shown that using human-annotated rationales, in the form of highlights that explain why a text has a specific label, can have positive effects on model generalisability. In this work, we ask to what extent attention regularisation with human-annotated rationales improve model performance and model robustness, as well as susceptibility to spurious correlations. We compare regularisation on human rationales with randomly selected tokens, a baseline which has hitherto remained unexplored. Our results suggest that often, attention regularisation with randomly selected tokens yields similar improvements to attention regularisation with human-annotated rationales. Nevertheless, we find that human-annotated rationales surpass randomly selected tokens when it comes to reducing model sensitivity to strong spurious correlations.

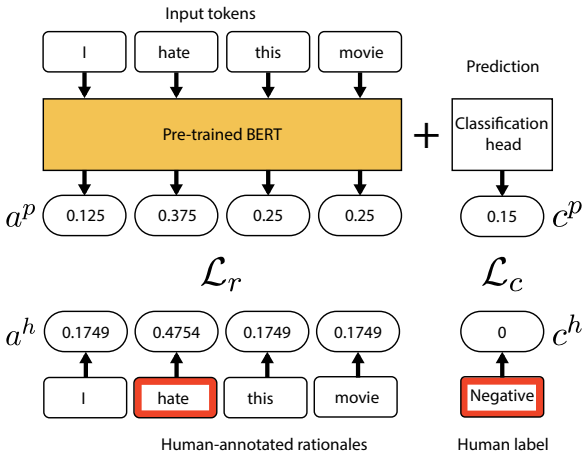


Figure 1: An image of attention regularisation using human-annotated rationales. Boxes with a red border represent human-annotated rationales and labels. The regularisation loss function ($\mathcal{L}_r$) measures the alignment between the attention pattern extracted from the model (a^p) and the reference attention pattern (a^h). The classification loss function ($\mathcal{L}_c$) measures the difference between the predicted (c^p) and the human (c^h) labels.

1 Introduction

The attention mechanism, as introduced by Vaswani et al. (2017), has become a key concept in NLP (Zhang and Kim, 2023). Attention regularisation aims to supervise the learned attention patterns in machine learning models equipped with the attention mechanism. Previous studies have shown that regularizing the attention patterns in Transformer models like BERT (Devlin et al., 2019) can accelerate convergence using less data (Deshpande and Narasimhan, 2020; Xia et al., 2021), improve downstream task performance and robustness (Joshi et al., 2022; Pruthi et al., 2022; Stacey et al., 2022), and reduce model bias (Mathew et al., 2021; Attanasio et al., 2022).

One way to regularise attention patterns is using human-annotated rationales (Zaidan et al., 2007; Herrewijnen et al., 2024). Human-annotated rationales are natural language explanations that explain why a human annotator labelled an instance with a label. For instance, a movie review can be labelled positive or negative, and specific words in the text may explain the chosen label (see Figure 1). The goal of using human-annotated rationales in attention regularisation is to encourage the model to rely on input cues that resemble the cues that humans would use to perform the task, thereby hoping to improve generalisability (Stacey et al., 2022) and reducing reliance on spurious correlations (Mathew et al., 2021).

While human-annotated rationales are a promising resource for improving NLP models, it remains unclear whether above performance improvements are caused by human-annotated rationales, as some related work has also reported improvements in downstream task performance when regularizing the model to distribute attention across all tokens

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 720–751

(Attanasio et al., 2022), or to disregard tokens that should be informative for performing the task (Pruthi et al., 2020; Pieke, 2023). Moreover, recent work has disputed whether attention regularisation truly changes the decision-making process of a trained model (Ferreira et al., 2025).

In this work, we aim to better understand the effects on the decision-making process of a model when applying attention regularisation using human-annotated rationales. To this end, we perform various experiments:[1]

- We compare attention regularisation using human-annotated rationales with multiple baselines (Section 3.1), finding that attention regularisation using randomly selected tokens can also improve downstream task performance. We hypothesise that this is due to the additional noise introduced during attention regularisation.
- We apply SHAP (Lundberg and Lee, 2017) to generate post-hoc explanations for models trained in different setups (Section 3.4). Our results show that changes in attention patterns are not echoed by generated SHAP explanations, further suggesting that attention regularisation does not (directly) change the model's decision process.
- To further understand our findings, we apply our analyses to a synthetically modified dataset containing spurious correlations (Section 3.2.2). Our findings indicate that using human-annotated rationales in attention regularisation can reduce the model's sensitivity to spurious correlations.

2 Preliminaries

In the following, we will discuss methods and parameters used in attention regularisation. We specifically focus on BERT (Devlin et al., 2019), as this model is often used in attention regularisation studies. Following Stacey et al. (2022), we also apply our methods to DeBERTa (He et al., 2021).

Attention regularisation consists of two objectives; first, the model learns to perform a downstream task, like sentiment analysis. The model learns to predict the most likely C label(s), thus outputting a C-dimensional vector. The goal is minimizing the loss $\mathcal{L}_c$ (e.g., cross-entropy loss) between predicted labels $\boldsymbol{c}^p$ and (human) reference labels $\boldsymbol{c}^h$. The second objective is attention regularisation. Here, the goal is to align $\boldsymbol{a}^p$, which are the attention patterns extracted from the model, with some reference attention pattern $\boldsymbol{a}^h$, by minimizing the regularisation loss $\mathcal{L}_r$. Then, $\mathcal{L}_c$ and $\mathcal{L}_r$ are combined using regularisation weight λ.

2.1 Extracting attention patterns from BERT

Let n be the number of tokens in an input text. The raw attention weights extracted from BERT consist of the attention values of every token in the input to all tokens, per layer and head. To construct attention pattern $\boldsymbol{a}^p$, the raw attention weights are aggregated to a sequence of n values, where every value represents the attention weight to a token in the input text ($\boldsymbol{a}^p = [a_0, a_1..., a_{n-1}, a_n]$) (see Figure 1). For downstream classification tasks, BERT uses the special classification token ([CLS]) to construct a sequence representation (Devlin et al., 2019). Therefore, related work usually regularises the attention patterns of the [CLS] token (Ferreira et al., 2025; Pruthi et al., 2022; Stacey et al., 2022), as the expectation is that the weights for this token change most after the model is fine-tuned on a downstream task. There is less consensus about which heads and layers to regularise; some related work regularises all attention heads and layers (Jayaram and Allaway, 2021), a fraction of heads in every layer (Deshpande and Narasimhan, 2020; Stacey et al., 2022), or only the final layer of the BERT model (Mathew et al., 2021; Jayaram and Allaway, 2021; Joshi et al., 2022; Ferreira et al., 2025). Moreover, Stacey et al. (2022) experiment with regularizing various subsets of attention heads, and find that regularizing the top 3 heads yields the greatest improvements on classification accuracy.

2.2 Reference attention patterns

Related work has used various types of reference attention patterns in attention regularisation. Some work employs attention regularisation to prevent the attention mechanism from paying attention to specific tokens (Attanasio et al., 2022). Alternatively, attention patterns extracted from large pretrained models (e.g., attend to the next or previous token) (Deshpande and Narasimhan, 2020), word similarity scores (Xia et al., 2021), and human-annotated word importance scores (Jayaram and Allaway, 2021; Pruthi et al., 2022; Mathew et al., 2021) are used as reference attention patterns.

[1]The code for our experiments is available at https://github.com/UtrechtUniversity/attention-regularisation-with-rationales

Human-annotated rationales as reference attention pattern. In this work, we focus on using human-annotated rationales as a reference attention pattern. Specifically, we use highlighted words in a text, that explain why a human annotator labelled an instance with a label. Human-annotated rationales have been successfully used for attention regularisation; Stacey et al. (2022) apply attention regularisation using human-annotated rationales to improve the robustness of their natural language inference models. Mathew et al. (2021) apply attention regularisation with human rationales to reduce unintended bias for models trained on the HateXplain dataset. Pruthi et al. (2022) use attention regularisation to improve performance on classification and question answering tasks. Lastly, Jayaram and Allaway (2021) use attention regularisation to improve few-shot classification performance.

Representing human-annotated rationales. Human-annotated rationales are often collected as binary values (e.g., a token is part of a rationale or not) (Herrewijnen et al., 2024). As attention patterns follow a probabilistic distribution, human-annotated rationales cannot be directly used as a reference attention pattern. To convert human-annotated rationales to a reference attention pattern $\boldsymbol{a}^h$, related work often normalises binary rationale values to continuous values that sum to 1 (Jayaram and Allaway, 2021; Stacey et al., 2022; Pruthi et al., 2022) or take the softmax (Mathew et al., 2021; Joshi et al., 2022; Ferreira et al., 2025) of binary rationale values. Note that for long input texts, the values per token may become very small.

2.3 Regularisation loss functions

To encourage $\boldsymbol{a}^p$ to align with $\boldsymbol{a}^h$, the loss term $\mathcal{L}_r$ is introduced during training. Various loss functions have been defined and are reviewed below.

Squared and Absolute Error. Various versions of error measures have been applied to calculate $\mathcal{L}_r$, including Mean Squared Error ($\mathcal{L}_{MSE}$) (Deshpande and Narasimhan, 2020; Joshi et al., 2022; Jayaram and Allaway, 2021), Summed Squared Error (Stacey et al., 2022), and Mean Absolute Error ($\mathcal{L}_{MAE}$) (Joshi et al., 2022).

Kullback–Leibler Divergence. Pruthi et al. (2022) and Joshi et al. (2022) use the Kullback–Leibler divergence ($\mathcal{L}_{KLD}$) to calculate $\mathcal{L}_r$, which measures how different two probability distributions are.

Order Loss. Joshi et al. (2022) introduce Order Loss ($\mathcal{L}_{OL}$), which aims to enforce that the weights for all rationale tokens are higher than the weights of all non-rationale tokens.

Attention Mass. We adapt the attention mass ($\mathcal{L}_{AM}$) loss function proposed by Pruthi et al. (2020) to sum the attention weights of all tokens that are not part of a rationale (eq. 1). This should ensure that the model pays the most attention to tokens that are part of a rationale.

$$\mathcal{L}_{AM} = \frac{1}{n} \sum_{\substack{t=1 \\ a_t^h \neq 1}}^{n} a_t^p \tag{1}$$

2.4 Regularisation weight

To apply attention regularisation, $\mathcal{L}_c$ and $\mathcal{L}_r$ need to be combined. For losses with very different magnitudes, simply summing the two losses may cause the optimiser to only focus on one of the two. To balance the two losses, an additional weight λ is added, such that the final loss function is $\mathcal{L} = \mathcal{L}_c + \lambda\mathcal{L}_r$. Some related work determines the best λ through hyperparameter tuning (Jayaram and Allaway, 2021; Deshpande and Narasimhan, 2020; Stacey et al., 2022; Pruthi et al., 2020). Alternatively, the regularisation weight remains unadjusted ($\lambda = 1$) (Stacey et al., 2022; Ferreira et al., 2025; Joshi et al., 2022).

Only a few studies have focused on comparing attention regularisation approaches; Joshi et al. (2022) explore various forms of explanation regularisation, where a model is regularised to yield explanations that align with human-annotated rationales. Next to gradient-based and perturbation-based explanation methods, they also include attention patterns as explanations. Moreover, they compare various regularisation loss functions (with $\lambda = 1$), and only report significant performance gains on OOD data for $\mathcal{L}_{MAE}$. Ferreira et al. (2025) compare attention regularisation for attention patterns extracted from the final layer, and patterns extracted using attention rollout (Abnar and Zuidema, 2020). Furthermore, they apply post-hoc explainability techniques to show that regularised models do not rely more on human-annotated rationale tokens than models without regularisation.

Complementary to existing work, we take a more systematic approach to determine the effects of attention regularisation; we apply our analysis to various baselines and datasets, including a novel

synthetic dataset. Moreover, we perform more elaborate hyper-parameter tuning of the λ value.

3 Methods

Determining which tokens a model used to perform a task is non-trivial. Therefore, we perform multiple experiments to measure the effects of attention regularisation using human-annotated rationales on the model's decision-making process. We train various models using three different types of reference attention patterns (Section 3.1) and a baseline model trained without any attention regularisation (None). For regularisation, we use the the regularisation loss functions discussed in Section 2.3.

We compare downstream task performance of our models on in-domain (ID) and out-of-domain (OOD) datasets (Section 3.2). Moreover, we apply our analysis to a dataset with synthetically introduced spurious correlations (Section 3.2.2). Finally, we analyse how well post-hoc explanations using SHAP (Lundberg and Lee, 2017) align with human-annotated rationales (Section 3.4).

Following related work (Pruthi et al., 2022; Jayaram and Allaway, 2021), we train our models on subsets of the data to determine the effect of attention regularisation in scarce data settings. See Appendix A and B for further training details.

3.1 Reference attention patterns

To compare our attention regularisation approaches, we use various reference attention patterns to regularise our models. Firstly, we use softmaxed binary human-annotated rationales (Section 2.2) as a reference attention pattern. As a second reference pattern, we use randomly selected tokens. For every instance in the training set, we note the number of human-annotated rationale tokens as r. Then, we randomly select r tokens from the input text as a reference attention pattern. If possible,[2] we only select tokens that are not human-annotated rationale tokens. We hypothesise that attention regularisation using human-annotated rationale tokens will outperform models trained with other types of reference attention patterns, as the latter should be less informative for models in performing the task.

Lastly, we apply Entropy-based Attention Regularisation ($\mathcal{L}_{EAR}$) by Attanasio et al. (2022), which regularises the attention mechanism to distribute attention across all tokens in the input. More specifically, $\mathcal{L}_{EAR}$ calculates the entropy of the softmaxed attention weights ($\boldsymbol{a}^p$) as a loss function:

$$\mathcal{L}_{EAR} = -\frac{1}{n}\sum_{t=1}^{n} a_t^p \log(a_t^p) \quad (2)$$

3.2 Tasks and Datasets

We train our models on three tasks using four empirical datasets, and on a dataset with synthetic spurious correlations.

3.2.1 Empirical datasets

For three tasks with various degrees of complexity, we train our models on in-domain (ID) data, and evaluate on a ID test set and multiple out-of-domain (OOD) test sets. By evaluating on OOD data, we aim to determine how well the model generalises to unseen domains. If a model performs well on both ID and OOD data, this suggests that the model learned to solve the task without relying on dataset-specific artifacts.

Our first task is sentiment analysis, for which we use two datasets for training. The first is the IMDb dataset (Zaidan et al., 2007), which contains movie reviews enriched with rationales. The second is the SST (Socher et al., 2013), which contains much shorter snippets from movie reviews. For OOD evaluation, we use DynaSent (Potts et al., 2021), Yelp (Zhang et al., 2015) and SFU reviews (Taboada and Grieve, 2004). The second task is hate speech detection, using HateXplain (Mathew et al., 2021) as an ID dataset. Hate speech detection is considered a complex task that is sensitive to bias (Mathew et al., 2021). As OOD datasets, we use Dynahate (Vidgen et al., 2021), Ethos (Mollas et al., 2022), and the White Supremacist Forum (WSF) (de Gibert et al., 2018) datasets. The third task is stance detection. Here, the task is to determine whether the text is in favour or against a specific topic. The stance of the text may be different per topic (e.g., in favour of Trump and against Biden), making the task non-trivial. We use the VAST (Allaway and McKeown, 2020) dataset for ID training, and the SemEval-2016 Stance (Mohammad et al., 2016) and PStance (Li et al., 2021) datasets for OOD evaluation.

To limit computational costs, we take a random subset of 1000 examples for our OOD evaluation. We further describe all tasks and datasets in Appendix D.1. See Table 4 (Appendix) for example annotations for every ID dataset.

[2] For HateXplain and SST, texts may contain a majority of human-annotated rationale tokens. If this is the case, we take all non-rationale tokens as the randomly selected reference attention pattern. If all tokens in the text are part of a human-annotated rationale, we randomly select $max(\frac{r}{2}, 1)$ tokens.

3.2.2 Synthetic spurious correlations

For the above empirical datasets, there may be multiple (sets of) tokens on which the model can rely to perform the task. To guide the model to rely on specific tokens to perform the task, we modify the IMDb dataset by adding spurious correlations. We expect the model to rely on these spurious correlations when performing the task.

First, we sample 500 examples from the dataset, of which we modify a percentage (25%, 50%, or 75%) by replacing all dots and question marks (. and ?) with an exclamation point (!) in positive examples, and replace all dots and exclamation points (. and !) with a question mark (?) in negative examples. This approach introduces spurious correlations to the training dataset, without (likely) changing the actual sentiment of the texts.

We then test the model on three test sets: the first contains instances with spurious correlations, for which we expect the model to reach near-perfect performance. The second test contains flipped spurious correlations, meaning that they are opposite to the spurious correlations (e.g., replacing dots and exclamation marks with a question mark point in positive examples), on which models that rely on the spurious correlations during training should fail. The third test set contains unmodified examples, on which such models should perform worse than on the spurious correlations test set, but better than on the flipped spurious correlations test set.

By applying attention regularisation, we aim to prevent the model from relying on spurious correlations. When this is the case, this should be visible in improved performance on the test set with flipped spurious correlations and the unmodified test set. Note that we exclude spurious correlation tokens (i.e., ! and ?) from the randomly selected tokens baseline (see Section 3.1).

3.3 Evaluation metrics

To evaluate the effectiveness of attention regularisation using human-annotated rationales, we apply two evaluation metrics. Firstly, to determine how well the model can perform the classification task, we report the macro average F1-score for the model's classification output. Secondly, to measure the intended change in attention patterns, i.e., pay more attention to human-annotated rationale tokens, we calculate the alignment between $\boldsymbol{a}^p$ and $\boldsymbol{a}^h$. Following Ferreira et al. (2025) and Fomicheva et al. (2021), we use the area under the receiver operating characteristic curve (AUC) to calculate alignment. AUC measures how likely a model is to rank a randomly selected rationale token above a randomly selected non-rationale token.

3.4 SHAP explanations

To estimate whether the model relies on human-annotated rationale tokens, we generate explanations using SHAP[3] (Lundberg and Lee, 2017). We chose SHAP as it allows for efficient generation of post-hoc and model-agnostic explanations that align well with human intuition (Retzlaff et al., 2024). Moreover, compared to LIME (Ribeiro et al., 2016), SHAP offers more robust and global insights (Roshinta and Gábor, 2024). For 100 randomly selected instances[4] from the test set, we use SHAP to calculate importance values for every token in every instance for the predicted label. We apply softmax to map the calculated importance values to a probability distribution $\boldsymbol{s}^p$. We further apply AUC (Section 3.3) to calculate the alignment between (binary) human-annotated rationales $\boldsymbol{a}^h$ and $\boldsymbol{s}^p$. Here, we expect that explanations for models regularised using human-annotated rationale tokens will align better with human-annotated rationales than models regularised using randomly selected tokens.

4 Results

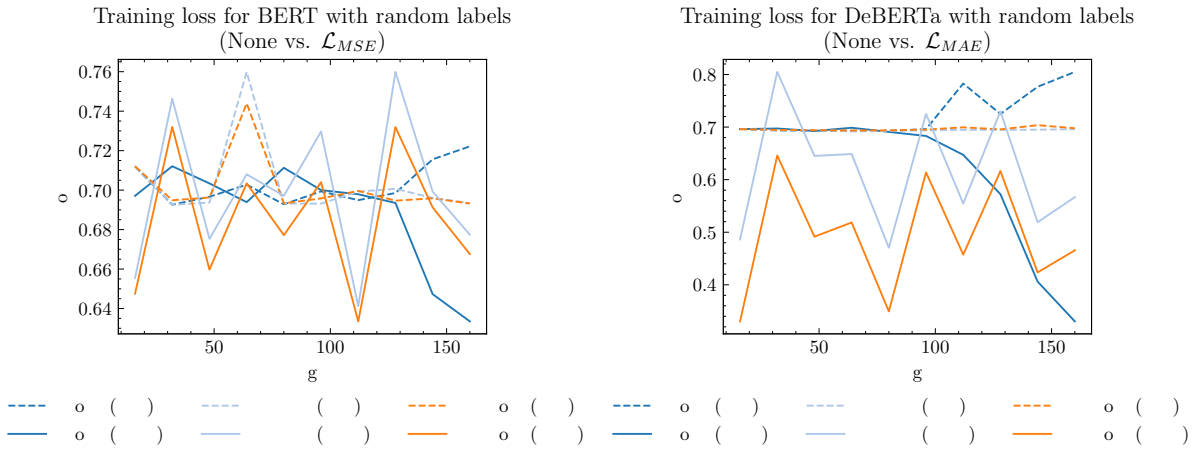
We now analyse how attention regularisation using human-annotated rationales affects classification performance. Similarly to Stacey et al. (2022), improvement gains are more visible in BERT than in DeBERTa models. Nevertheless, our results suggest that attention regularisation can improve classification performance in both models.

In general, there is no regularisation loss function ($\mathcal{L}_r$) that outperforms all other $\mathcal{L}_r$ functions. In Figure 4 (Appendix), we show the Spearman rank correlations between various metrics and hyperparameters. The size of the training dataset (DS) correlates positively with F1-scores for ID and OOD, showing that the models generalise better when more data is available during training.

Attention regularisation improves ID and OOD performance. Similarly to related work (e.g., Pruthi et al. (2022)) attention regularisation improves ID and OOD classification performance most when the training dataset is small (e.g., 100

[3] See Appendix C for more details.

[4] Due to the small test set, this number is 21 for VAST.



(a) BERT regularised with $\mathcal{L}_{MSE}$.

(b) DeBERTa regularised with $\mathcal{L}_{MAE}$

Figure 2: Training loss for BERT and DeBERTa models trained on randomly shuffled labels. We train without attention regularisation (None), and using $\mathcal{L}_{MSE}$ (BERT) and $\mathcal{L}_{MAE}$ (DeBERTa) with human-annotated rationales (h) and randomly selected tokens (r). We show the training (solid lines), and validation (dashed lines) loss. Models trained without attention regularisation (None) begin to overfit around step 100 (BERT) and 70 (DeBERTa), showing that attention regularisation hinders overfitting.

or 200 examples). Regularisation using $\mathcal{L}_{OL}$ negatively impacts performance for DeBERTa, where the model does not converge (Table 11-14, Appendix). Moreover, for all datasets except HateXplain, attention regularisation yields improvements in OOD performance.[5] See Appendices E and G for ID and OOD performance of all setups.

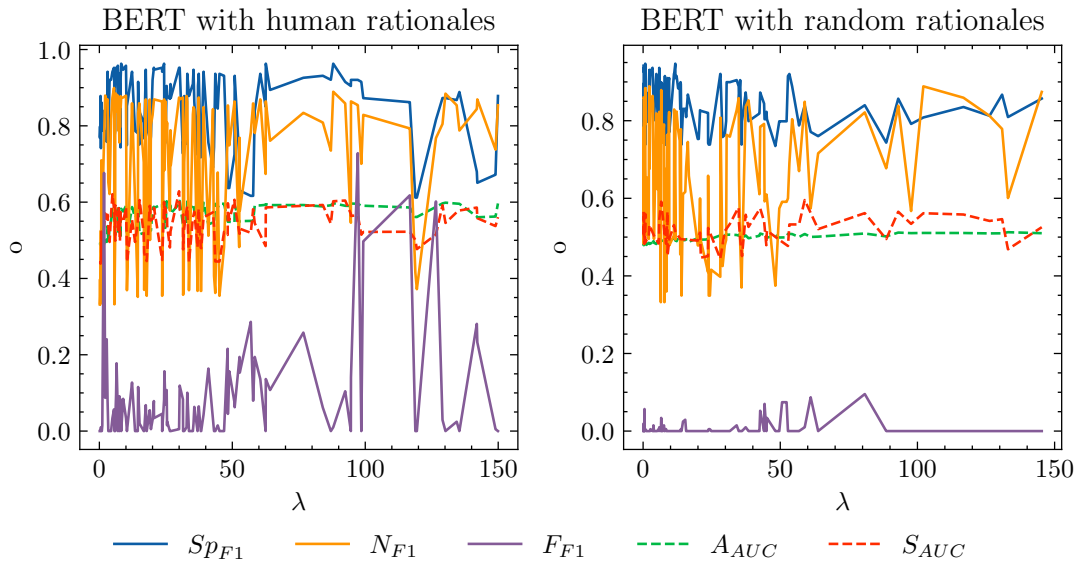
Attention regularisation does not change SHAP explanations. For the datasets used in this work, attention regularisation appears to have little effect on generated SHAP explanations. AUC scores for attention patterns (A_{AUC}) do not correlate with AUC scores for SHAP values (S_{AUC}) (Figure 4, Appendix), indicating that SHAP explanations do not echo changes in attention patterns. Nevertheless, we observe that scores for ID_{F1}, OOD_{F1}, and S_{AUC} correlate positively, suggesting that when models perform the task accurately, the generated SHAP explanations align better with human-annotated rationales.

Random tokens can be useful reference attention patterns. Surprisingly, the above performance improvements can be achieved by regularising using either human-annotated rationale tokens, or randomly selected tokens. Regularising using a uniformly distributed attention pattern ($\mathcal{L}_{EAR}$) performs similarly to the model without attention regularisation (Appendices E and G). We surmise that attention regularisation might introduce additional noise during training, thus helping to prevent the model from overfitting on the ID data. To test this hypothesis, we train our models on a dataset with shuffled labels, and compare training losses over training steps. Our expectation is that the model trained with attention regularisation will be slower to converge due to the additional noise. In Figure 2, we show the training loss over training steps for models trained without attention regularisation, and using $\mathcal{L}_{MSE}$ for BERT and $\mathcal{L}_{MAE}$ for DeBERTa.[6] As expected, the model trained without attention regularisation (None) begins to overfit after a number of steps, but the regularised models do not converge. Moreover, the loss patterns for models regularised with human-annotated rationale tokens (human) and randomly selected tokens (random) are quite alike.

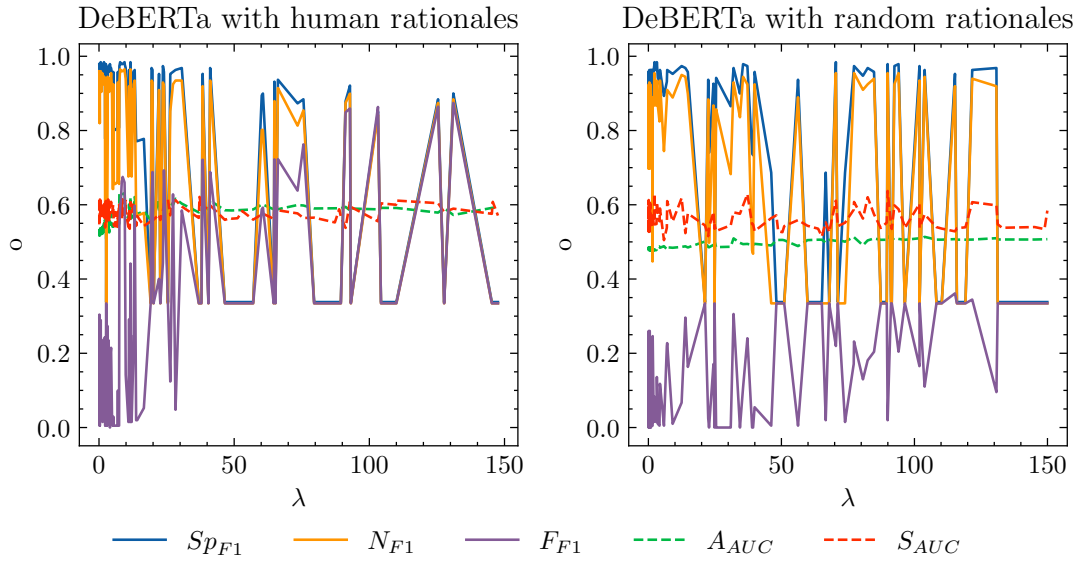
Attention regularisation with human-annotated rationales lowers sensitivity to spurious correlations. Applying attention regularisation on the dataset with synthetically introduced spurious correlations (Section 3.2.2) shows that human-annotated rationale tokens yield greater benefits than randomly selected tokens in attention regularisation (Table 1). As expected, the model without attention regularisation performs well on a dataset

[5]Note that for DeBERTa trained on all examples from the IMDb set, attention regularisation negatively impacts OOD performance (Table 21).

[6]We choose these attention regularisation methods because of the performance improvements on the IMDB dataset, see Appendices E and G.



(a) BERT regularised using $\mathcal{L}_{MSE}$.



(b) DeBERTa regularised using $\mathcal{L}_{MAE}$.

Figure 3: Performance and λ of models trained on 500 examples with 50% spurious correlations. The BERT and DeBERTa models are trained with attention regularisation using $\mathcal{L}_{MSE}$ and $\mathcal{L}_{MAE}$, respectively. Sp_{F1} denotes the macro average F1-score for the test set with spurious correlations. A_{AUC} denotes the AUC for attention patterns and human-annotated rationale tokens, and S_{AUC} for SHAP scores. N_{F1} denotes the F1-score for the test set without synthetically introduced spurious correlations. Finally, F_{F1} denotes the F1-score for the test set with flipped spurious correlations. Higher λ values clearly improve F_{F1} when the model is regularised using human-annotated rationale tokens.

$SC_{\%}$		25%			50%			75%		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	Sp_{F1}	F_{F1}	N_{F1}	Sp_{F1}	F_{F1}	N_{F1}	Sp_{F1}	F_{F1}	N_{F1}
None		1.00	.07	.89	1.00	.00	.83	1.00	.00	.55
$\mathcal{L}_{EAR}$		1.00	.05	.78	1.00	.00	.78	1.00	.00	.57
$\mathcal{L}_{AM}$	h	.99	**.23**	.88	1.00	.00	.79	1.00	.00	.64
	r	1.00	.11	.91	1.00	.00	.71	1.00	.00	.59
$\mathcal{L}_{KLD}$	h	1.00	.11	.91	1.00	**.06**	.88	1.00	.00	**.82**
	r	1.00	.08	.89	1.00	.00	.73	1.00	.00	.62
$\mathcal{L}_{MAE}$	h	1.00	.10	.90	1.00	**.05**	.87	1.00	.00	**.78**
	r	1.00	.08	.90	1.00	.00	.86	1.00	.00	.43
$\mathcal{L}_{MSE}$	h	1.00	**.27**	.90	1.00	**.07**	.86	1.00	.01	**.83**
	r	1.00	.09	.86	1.00	.00	.80	1.00	.00	.57
$\mathcal{L}_{OL}$	h	1.00	.07	.87	1.00	.01	.75	1.00	.00	.76
	r	1.00	.07	.84	1.00	.02	.85	1.00	.00	.77

(a) Averaged results for BERT models trained on examples containing spurious correlations.

$SC_{\%}$		25%			50%			75%		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	Sp_{F1}	F_{F1}	N_{F1}	Sp_{F1}	F_{F1}	N_{F1}	Sp_{F1}	F_{F1}	N_{F1}
None		.95	.51	.89	1.00	.16	.90	1.00	.05	.88
$\mathcal{L}_{EAR}$		.95	.51	.89	1.00	.16	.90	1.00	.05	.88
$\mathcal{L}_{AM}$	h	.99	.47	.94	.99	.34	.94	1.00	.08	.93
	r	.99	.43	.94	.99	.31	.93	1.00	.08	.93
$\mathcal{L}_{KLD}$	h	.97	.65	.94	1.00	.17	.93	1.00	.06	.93
	r	.99	.56	.94	1.00	.13	.94	1.00	.06	.93
$\mathcal{L}_{MAE}$	h	.98	**.73**	.89	1.00	**.30**	.95	.99	**.37**	.94
	r	.98	.52	.94	1.00	.21	.94	1.00	.20	.93
$\mathcal{L}_{MSE}$	h	.98	.55	.94	1.00	.17	.85	1.00	.06	.86
	r	.98	.55	.94	1.00	.17	.85	1.00	.06	.86
$\mathcal{L}_{OL}$	h	.97	.71	.94	.99	.45	.92	.85	**.45**	.69
	r	.97	.74	.93	.99	.45	.90	.77	.31	.73

(b) Averaged results for DeBERTa models trained on examples containing spurious correlations.

Table 1: Averaged results (3 random seeds) for BERT and DeBERTa models trained on 500 examples from the IMDb dataset with spurious correlations. $SC_{\%}$ denotes the percentage of synthetic spurious correlations in the training set. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r). Sp_{F1} refers to the F1-score on the dataset with spurious correlations, F_{F1} to the F1-score on the dataset with flipped correlations, and N_{F1} to the F1-score on the dataset without spurious correlations. **Bold** values denote setups where attention regularisation using human-annotated rationale tokens outperform randomly selected tokens. We report the standard deviation values in Appendix F.

with spurious correlations, but less well on the dataset without spurious correlations, and poorly on a dataset with flipped correlations (Table 1). This indicates that these models rely heavily on spurious correlations, especially when a greater percentage of training examples contains spurious signal. Applying attention regularisation improves performance on the test set with flipped spurious correlations (F_{F1}) and the unmodified test set (N_{F1}), especially when using human-annotated rationales. We note that DeBERTa is less susceptible to learning spurious correlations (Table 3b).

Moreover, we find further evidence that human-annotated rationales are more useful in reducing sensitivity to spurious correlations than randomly selected tokens; for higher values of λ, models regularised using human-annotated rationales perform better on the test set with flipped spurious correlations (F_{F1}). This is shown in Figure 3, where we show performance of models trained on 500 examples from IMDb with 50% spurious correlations for various values of λ.

The regularisation weight should be finetuned per task and dataset. The regularisation loss $\mathcal{L}_r$ often has a different scale than the classification loss $\mathcal{L}_c$, due to the output dimensionality (C or n). Moreover, the number of tokens and rationale tokens in the training dataset influence the scale of $\mathcal{L}_r$. Table 6b (Appendix) shows the $\mathcal{L}_c$ across different loss functions and datasets extracted from an untrained BERT model. We note that some loss functions are affected by the token and rationale length of training examples. The optimal hyper-parameters found through hyper-parameter tuning (Table 2 and 3, Appendix) underline this. The regularisation weight (λ) differs greatly across functions for $\mathcal{L}_r$ and dataset sizes. Finally, though it is dependent on both loss function and data size, our results suggest that increasing the value for λ can render human-annotated rationales more effective.

5 Conclusion and Discussion

In this work, we systematically investigate how attention regularisation using human-annotated rationales impacts classification models. Our results suggest that attention regularisation can improve classification performance on both in-domain and out-of-domain settings, especially for scarce datasets. However, to our surprise, regularisation using randomly selected token often yielded similar performance improvements.

Post-hoc explanations generated using SHAP do not resemble human-annotated rationales more after applying attention regularisation with human-annotated rationale tokens. This suggests that attention regularisation only subtly changes the trained model's decision-making process, and/or that SHAP is unable to detect nuanced changes in the decision-making process of the model. Ferreira et al. (2025) report similar results for other types of post-hoc explanations, further indicating that it may be difficult to detect changes in the decision-making process of regularised models.

Taken together, our results suggest that performance improvements of attention regularisation are mostly due to the additional noise introduced during training, thus hindering the model from overfitting (Bishop, 1995).

Nevertheless, we show that attention regularisation using human-annotated rationale tokens can reduce sensitivity to spurious correlations. More specifically, a higher λ value reduces the model's reliance on spurious correlations. While this is clearly visible in classification performance, the explanations generated using SHAP do not reflect this. Future work should further explore explainability methods and shifts in decision-making processes of machine learning models.

6 Limitations

A limitation of this work is that datasets used in attention regularisation experiments (Section 3.2) often consist of short texts with a relatively high number of rationale tokens (see Appendix D.4). For datasets with short texts, attention regularisation may have a limited effect on performance, as the model does not have to divide attention across many tokens. To test this hypothesis, we invite future work to explore attention regularisation for datasets with longer texts.

Using human-annotated rationales as a reference attention pattern has limitations (Tan, 2022); such rationales may not be exhaustive (Herrewijnen et al., 2024), meaning that other tokens in the text can also form part of plausible rationales. Furthermore, human annotators may disagree on which tokens should be annotated as rationales. Finally, tokens that are informative for a human may not be informative for a ML model, or vice versa.

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, Online. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using

Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

Chris M. Bishop. 1995. Training with Noise Is Equivalent to Tikhonov Regularization. *Neural Computation*, 7(1):108–116.

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, Online. Association for Computational Linguistics.

Benoit Courty, Victor Schmidt, Goyal-Kamal, Marion-Coutarel, Boris Feld, Jérémy Lecourt, LiamConnell, SabAmine, inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Amine Saboni, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, alencon, Michał Stęchły, Christian Bauer, Lucas-Otavio, JPW, and MinervaBooks. 2024. Mlco2/codecarbon: V2.4.1. Zenodo.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Ameet Deshpande and Karthik Narasimhan. 2020. Guiding Attention for Self-Supervised Learning with Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4676–4686, Online. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics.

Pedro Ferreira, Ivan Titov, and Wilker Aziz. 2025. Explanation Regularisation through the Lens of Attributions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6530–6551, Abu Dhabi, UAE. Association for Computational Linguistics.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Elize Herrewijnen, Dong Nguyen, Floris Bex, and Kees van Deemter. 2024. Human-Annotated Rationales and Explainable Text Classification: A Survey. *Frontiers in Artificial Intelligence*, 7.

Sahil Jayaram and Emily Allaway. 2021. Human Rationales as Attribution Priors for Explainable Stance Detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5540–5554, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A Large Dataset for Stance Detection in Political Domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings*

of the AAAI Conference on Artificial Intelligence, 35(17):14867–14875.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

Michael Pieke. 2023. Bert, but better: Improving robustness using human insights. Master's thesis, Utrecht University.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2388–2404, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Carl O Retzlaff, Alessa Angerschmid, Anna Saranti, David Schneeberger, Richard Roettger, Heimo Mueller, and Andreas Holzinger. 2024. Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cognitive Systems Research*, 86:101243.

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*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Trisna Ari Roshinta and Szűcs Gábor. 2024. A comparative study of lime and shap for enhancing trustworthiness and efficiency in explainable ai systems. In *2024 IEEE International Conference on Computing (ICOCO)*, pages 134–139.

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.

Maite Taboada and Jack Grieve. 2004. Analyzing Appraisal Automatically. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 158–161.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188.

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

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Tingyu Xia, Yue Wang, Yuan Tian, and Yi Chang. 2021. Using Prior Knowledge to Guide BERT's Attention in Semantic Textual Matching Tasks. In *Proceedings of the Web Conference 2021*, pages 2466–2475, Ljubljana Slovenia. ACM.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Nan Zhang and Junyeong Kim. 2023. A Survey on Attention mechanism in NLP. In *2023 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–4.

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

A Training details

Following previous work (Section 2), we extract $\boldsymbol{a}^p$ by taking the attention weights for the [CLS] token of the last layer of the model, averaged across heads.[7] We fine-tune the bert-base-uncased[8] model for 5 epochs and a batch size of 16. For every task, we train our models on 100, 200, and 500 randomly sampled examples, and on the full dataset. To mitigate the effect of randomness (Dror et al., 2019), we train our models using the found hyper-parameters on three random seeds, and report results averaged over these three models.

Hyper-parameters. To find the optimal hyper-parameters for model training, we first perform hyper-parameter tuning using the transformers library.[9] We use Optuna (Akiba et al., 2019) to search for the optimal learning rate and λ by maximizing the macro average F1-score for the classification task on the development set, across 20 trials. The search space for the learning rate lies between 1e-8 and 1e-4, and for λ between 0.1 and 100. We present hyper-parameters found through hyper-parameter search in Table 2 and 3.

Carbon footprint. We train our models using a system with two NVIDIA RTX A6000 cards for 2522 cumulative hours. Using the CodeCarbon library (Courty et al., 2024), we estimate the carbon emissions of this work to be 3.08 $kgCO_2eq$.

B Regularisation weight

We expect that there is a difference in scale between functions for $\mathcal{L}_r$ (which we confirm in Section 4). To understand the effect of λ on training results, e.g., should the optimiser prioritise $\mathcal{L}_r$ over $\mathcal{L}_c$, we first balance $\mathcal{L}_c$ and $\mathcal{L}_r$. To do this, we calculate the initial loss for $\mathcal{L}_r$ and $\mathcal{L}_c$ for an untrained model. By dividing $\mathcal{L}_c$ by $\mathcal{L}_r$, we find the balance modifier b. Using b, the final loss function used for training is as follows:

$$\mathcal{L} = \mathcal{L}_c + \lambda b \mathcal{L}_r \tag{3}$$

Using this approach, λ accurately represents the weight of $\mathcal{L}_r$ relative to $\mathcal{L}_c$, while limiting the hyper-parameter search space for λ.

[7] We also use attention rollout and the average of the top 3 heads in BERT, but find this does not significantly change results.

[8] https://huggingface.co/google-bert/bert-base-uncased

[9] https://github.com/huggingface/transformers

C SHAP explanations

We use the python SHAP library[10], version 0.47.2, to generate SHAP explanations. Moreover, we use the 'partition' algorithm to estimate the Shapley values.

D Datasets

In the following, we describe the datasets used in this study. See Table 5 for the number of instances per dataset.

D.1 Description of tasks and datasets

Sentiment Analysis (IMDb and SST). For sentiment analysis, we use two datasets. The first is the Internet Movie Database (IMDb) dataset enriched with rationales (Zaidan et al., 2007), which contains English movie reviews with positive or negative sentiment labels and human-annotated snippet rationales explaining the labels. Human-annotated rationales are collected by instructing annotators to justify their decision by highlighting the most important words and phrases in the review.

The second dataset is the Stanford Sentiment Treebank (SST) (Socher et al., 2013), which contains snippets from movie reviews with positive or negative sentiment labels at every syntactic tree node. We use the heuristic algorithm by Carton et al. (2020) to construct human-annotated rationales. Compared to the IMDb dataset, the SST dataset contains rather short texts with a high percentage of rationales per text (Table 6a).

As OOD datasets, we use the DynaSent (Potts et al., 2021), Yelp (Zhang et al., 2015) and SFU review (Taboada and Grieve, 2004) datasets. DynaSent is a challenge dataset for sentiment classification and contains two rounds of challenge sentences. The first round contains naturally occurring sentences that are challenging for machine learning models to classify. The second round contains sentences crafted by humans to 'fool' machine learning models. The Yelp dataset contains user reviews for local businesses annotated with positive or negative sentiment labels. The SFU review dataset consists of consumer reviews for products from Epinions.com, labelled as recommended (positive) or not recommended (negative).

Hate speech detection (HateXplain). The HateXplain dataset (Mathew et al., 2021) contains Twitter and Gap posts, annotated with hate speech,

[10] https://pypi.org/project/shap/

DS		100		200		500		1472	
	b	lr	λ	lr	λ	lr	λ	lr	λ
None		3×10^{-5}		6×10^{-5}		10×10^{-5}		5×10^{-5}	
$\mathcal{L}_{EAR}$	-0.05	2×10^{-5}	22.79	7×10^{-5}	1.03	7×10^{-5}	4.66	2×10^{-5}	3.06
$\mathcal{L}_{AM}$	0.42	3×10^{-5}	1.75	5×10^{-6}	1.95	7×10^{-5}	1.42	9×10^{-5}	1.65
$\mathcal{L}_{KLD}$	748.82	3×10^{-5}	7.56	6×10^{-5}	1.01	7×10^{-5}	0.28	4×10^{-5}	11.20
$\mathcal{L}_{MAE}$	6332.53	6×10^{-5}	1.13	7×10^{-5}	0.13	2×10^{-5}	25.51	2×10^{-5}	0.61
$\mathcal{L}_{MSE}$	54 185 485.09	10×10^{-5}	61.40	1×10^{-5}	8.60	8×10^{-5}	0.34	9×10^{-5}	0.10
$\mathcal{L}_{OL}$	16.59	1×10^{-5}	0.85	7×10^{-5}	0.38	4×10^{-5}	3.84	5×10^{-5}	0.20

(a) IMDb

DS		100		200		500		6541	
	b	lr	λ	lr	λ	lr	λ	lr	λ
None		1×10^{-7}		10×10^{-5}		4×10^{-5}		4×10^{-5}	
$\mathcal{L}_{EAR}$	0.00	3×10^{-5}	0.11	9×10^{-5}	0.67	3×10^{-5}	0.23	5×10^{-6}	8.31
$\mathcal{L}_{AM}$	1.82	4×10^{-5}	12.22	6×10^{-5}	11.95	10×10^{-5}	37.92	7×10^{-5}	7.53
$\mathcal{L}_{KLD}$	747.40	4×10^{-5}	12.77	1×10^{-8}	0.12	7×10^{-5}	1.29	3×10^{-5}	0.94
$\mathcal{L}_{MAE}$	318.29	2×10^{-5}	0.13	9×10^{-5}	0.37	3×10^{-5}	0.48	5×10^{-5}	4.21
$\mathcal{L}_{MSE}$	1 474 452.16	6×10^{-5}	8.40	6×10^{-5}	1.04	10×10^{-5}	4.51	4×10^{-5}	0.15
$\mathcal{L}_{OL}$	0.10	3×10^{-5}	0.23	5×10^{-5}	0.42	9×10^{-5}	0.15	5×10^{-5}	63.41

(b) SST

DS		100		200		500		14102	
	b	lr	λ	lr	λ	lr	λ	lr	λ
None		2×10^{-8}		1×10^{-8}		3×10^{-5}		1×10^{-5}	
$\mathcal{L}_{EAR}$	-0.05	1×10^{-5}	2.88	9×10^{-5}	6.32	6×10^{-8}	1.37	3×10^{-5}	0.34
$\mathcal{L}_{AM}$	1.22	3×10^{-6}	0.44	2×10^{-6}	13.72	3×10^{-8}	4.64	2×10^{-5}	20.33
$\mathcal{L}_{KLD}$	265.02	5×10^{-8}	15.55	9×10^{-5}	0.82	4×10^{-5}	10.64	2×10^{-5}	2.16
$\mathcal{L}_{MAE}$	28 098.90	5×10^{-7}	76.17	1×10^{-8}	0.10	7×10^{-5}	0.75	1×10^{-5}	2.30
$\mathcal{L}_{MSE}$	6 350 077.65	5×10^{-8}	39.73	9×10^{-8}	2.34	10×10^{-5}	0.11	2×10^{-5}	0.98
$\mathcal{L}_{OL}$	48.70	5×10^{-8}	27.23	5×10^{-7}	93.88	2×10^{-8}	1.39	3×10^{-5}	0.88

(c) HateXplain

DS		100		200		519	
	b	lr	λ	lr	λ	lr	λ
None		9×10^{-5}		6×10^{-6}		6×10^{-7}	
$\mathcal{L}_{EAR}$	-0.05	9×10^{-5}	1.05	7×10^{-6}	22.05	2×10^{-5}	0.27
$\mathcal{L}_{AM}$	9.14	5×10^{-5}	2.58	5×10^{-6}	95.59	4×10^{-5}	3.26
$\mathcal{L}_{KLD}$	17.20	3×10^{-5}	1.02	6×10^{-5}	1.00	2×10^{-6}	1.14
$\mathcal{L}_{MAE}$	651.34	3×10^{-5}	4.39	1×10^{-5}	1.68	1×10^{-6}	0.61
$\mathcal{L}_{MSE}$	30 041.67	9×10^{-5}	0.26	7×10^{-6}	0.46	2×10^{-6}	0.30
$\mathcal{L}_{OL}$	4.00	4×10^{-5}	6.08	4×10^{-5}	42.01	2×10^{-5}	1.96

(d) VAST

Table 2: Balance modifier (b), learning rate (lr), and λ found through hyper-parameter search for BERT.

DS		100		200		500		1472	
	b	lr	λ	lr	λ	lr	λ	lr	λ
None		4×10^{-5}		8×10^{-5}		4×10^{-5}		2×10^{-5}	
$\mathcal{L}_{EAR}$	0.00	8×10^{-5}	2.95	9×10^{-5}	16.59	4×10^{-5}	0.11	2×10^{-5}	0.10
$\mathcal{L}_{AM}$	795.39	3×10^{-5}	108.69	7×10^{-5}	124.73	3×10^{-5}	0.59	3×10^{-5}	27.11
$\mathcal{L}_{KLD}$	18.88	6×10^{-5}	29.82	7×10^{-5}	0.61	4×10^{-5}	1.75	7×10^{-6}	127.85
$\mathcal{L}_{MAE}$	3.45	7×10^{-5}	0.18	5×10^{-5}	6.79	2×10^{-5}	10.40	8×10^{-6}	4.88
$\mathcal{L}_{MSE}$	5 314 408.58	4×10^{-5}	18.07	6×10^{-5}	0.92	1×10^{-5}	10.90	8×10^{-6}	72.78
$\mathcal{L}_{OL}$	2529.40	4×10^{-5}	132.35	1×10^{-5}	13.70	4×10^{-5}	0.20	6×10^{-6}	0.13

(a) IMDb

DS		100		200		500		6541	
	b	lr	λ	lr	λ	lr	λ	lr	λ
None		3×10^{-6}		10×10^{-5}		8×10^{-5}		3×10^{-5}	
$\mathcal{L}_{EAR}$	-0.13	10×10^{-5}	7.66	6×10^{-5}	0.26	10×10^{-5}	0.11	4×10^{-5}	0.35
$\mathcal{L}_{AM}$	1.22	8×10^{-5}	0.35	8×10^{-5}	0.83	2×10^{-5}	14.59	1×10^{-5}	0.12
$\mathcal{L}_{KLD}$	495.26	5×10^{-6}	8.14	10×10^{-5}	0.60	3×10^{-5}	1.18	3×10^{-5}	1.15
$\mathcal{L}_{MAE}$	86.70	1×10^{-6}	0.11	9×10^{-5}	2.57	1×10^{-5}	0.39	2×10^{-5}	0.81
$\mathcal{L}_{MSE}$	149 395.21	10×10^{-5}	4.09	10×10^{-5}	0.52	2×10^{-5}	0.26	3×10^{-5}	0.51
$\mathcal{L}_{OL}$	25 287.37	4×10^{-8}	40.07	3×10^{-7}	3.91	7×10^{-5}	0.49	4×10^{-5}	15.32

(b) SST

DS		100		200		500		14102	
	b	lr	λ	lr	λ	lr	λ	lr	λ
None		10×10^{-7}		7×10^{-5}		4×10^{-5}		2×10^{-5}	
$\mathcal{L}_{EAR}$	-0.02	10×10^{-5}	0.86	7×10^{-5}	3.76	4×10^{-5}	3.26	9×10^{-5}	0.10
$\mathcal{L}_{AM}$	12.28	7×10^{-5}	1.39	1×10^{-7}	24.07	2×10^{-5}	21.57	8×10^{-6}	0.21
$\mathcal{L}_{KLD}$	3137.00	3×10^{-6}	44.32	4×10^{-8}	32.28	2×10^{-5}	0.70	1×10^{-6}	0.70
$\mathcal{L}_{MAE}$	443.54	9×10^{-5}	0.67	9×10^{-5}	0.12	6×10^{-5}	1.09	1×10^{-5}	0.71
$\mathcal{L}_{MSE}$	134 082.31	1×10^{-8}	6.28	8×10^{-5}	139.73	8×10^{-5}	0.23	2×10^{-6}	0.70
$\mathcal{L}_{OL}$	1269.52	4×10^{-7}	2.10	2×10^{-5}	12.15	5×10^{-5}	0.14	4×10^{-5}	45.53

(c) HateXplain

DS		100		200		519	
	b	lr	λ	lr	λ	lr	λ
None		7×10^{-5}		8×10^{-5}		4×10^{-5}	
$\mathcal{L}_{EAR}$	-0.15	4×10^{-5}	74.85	2×10^{-5}	11.32	6×10^{-5}	11.14
$\mathcal{L}_{AM}$	1.68	3×10^{-5}	19.28	2×10^{-5}	1.19	4×10^{-5}	12.74
$\mathcal{L}_{KLD}$	0.22	5×10^{-5}	0.48	1×10^{-5}	1.27	2×10^{-5}	1.20
$\mathcal{L}_{MAE}$	559.49	6×10^{-5}	0.85	4×10^{-5}	2.83	2×10^{-5}	0.42
$\mathcal{L}_{MSE}$	1 706 660.43	9×10^{-5}	4.78	4×10^{-5}	0.95	8×10^{-5}	0.12
$\mathcal{L}_{OL}$	96.73	1×10^{-7}	19.06	4×10^{-6}	142.30	6×10^{-5}	0.73

(d) VAST

Table 3: Balance modifier (b), learning rate (lr), and λ found through hyper-parameter search for DeBERTa.

abusive, or normal language. If the post contains hate speech or abusive language, the target community is also annotated. Every post is annotated by three annotators, and we combine the labels to construct a multi-label task. This means that posts can be both hate speech or abusive language, neither, or both. When a post is annotated as hate speech or abusive language, annotators are asked to select that words or phrases could be a potential reason for the given annotation.

Moreover, we combine all words annotated as rationales as our rationale mask. Note that neutral examples do not contain rationales. As OOD datasets, we use multiple hate speech datasets; the Dynahate dataset (Vidgen et al., 2021) contains hate speech texts created by annotators, which are perturbed by human annotators to 'flip the label' in three rounds. The Ethos dataset (Mollas et al., 2022), which contains posts from Reddit and YouTube annotated with hate speech labels. Lastly, the White Supremacist Forum dataset (de Gibert et al., 2018) consists of posts from the Stormfront forum, and is annotated with hate speech labels.

Stance detection (VAST). The VAried Stance Topics (VAST) dataset by Allaway and McKeown (2020) contains comments collected from The New York Times 'Room for Debate' section. For a given topic (e.g., marijuana), every instance is annotated pro, con, or neutral towards the topic. Jayaram and Allaway (2021) augment a non-neutral subset of VAST with human-annotated rationales. Annotators are asked to select the words that were most important to determine the stance towards the given topic. In this work, we only use the instances with human-annotated rationales, thus removing neutral instances. Note that this simplifies the task to binary classification (pro or con).

As OOD datasets, we use two stance detection datasets consisting of tweets. The first is the SemEval-2016 Stance dataset (Mohammad et al., 2016). This dataset consists of stances towards various topics, like presidential candidates and societal issues. Every instance is labelled in favour or against the topic. We remove neutral instances. The second dataset is the PStance dataset (Li et al., 2021), which contains tweets targeting presidential candidates. Every instance is labelled in favour or against the candidate. Following Jayaram and Allaway (2021), we modify the input text in all datasets to '[CLS] text [SEP] topic [SEP]'.

D.2 Preprocessing

We train our models on instances that include human-annotated rationales after tokenisation. We remove instances where all rationale tokens are truncated during tokenisation. Moreover, we remove instances where all tokens in the instance are rationales. For SST, this applies to 379 instances.

D.3 Examples of human-annotated rationales

See Table 4 for examples of rationales in the datasets used in this work.

D.4 Dataset statistics

In Table 6a we present the average number of tokens and rationale tokens per dataset used in this study. In Table 6b we present values for various $\mathcal{L}_r$ for an untrained BERT model.

E In-domain performance

In Tables 7-14, we present F1-scores on in-domain datasets, alignment of attention patterns with human-annotated rationale tokens (A_{AUC}), and alignment of extracted SHAP explanations with human-annotated rationale tokens (S_{AUC}) for our setups.

E.1 BERT

See Tables 7-10 for in domain performance for BERT models.

E.2 DeBERTa

See Tables 11-14 for in domain performance for DeBERTa models.

F Performance for dataset with spurious correlations

In Tables 15 and 16 we show the performance for BERT and DeBERTa models trained on various percentages (25%, 50% and 75%) of spurious correlations. All results are averaged over 3 random seeds.

G Out-of-domain performance

In Tables 17-24 we present F1-scores on out-of-domain datasets for our setups.

G.1 BERT

See Tables 17-20 for the OOD performance for BERT models.

IMDb	
Claire Danes, Giovanni Ribisi, and Omar Epps make a likable trio of protagonists, but they're just about the only palatable element of the mod squad, a lame-brained big-screen version of the 70s tv show. the story has all the originality of a block of wood (well, it would if you could decipher it), the characters are all blank slates. Scott silver's perfunctory action sequences are as cliched as they come. by sheer force of talent, the three actors wring marginal enjoyment from the proceedings whenever they're on screen. Still, the mod squad is just a second-rate action picture with a first-rate cast.	Negative
Hilarious, ultra-low budget comedy from film school dropout Kevin Smith chronicles a day in the life of two convenience store slackers (Brian o'Halloran and Jeff Anderson). they spend most of their day ignoring customers while discussing everything from fellatio to self-fulfillment. the premise is strictly sitcom and the photography is grainy as all get-out, but you could spend ten times the film's budget (a reported $ 27, 000) and still not get dialogue half as good as this. originally rated nc-17 for language. not recommended for viewers with sensitive ears.	Positive
SST	
I hate this movie	Negative
scores a few points for doing what it does with a dedicated and good-hearted professionalism.	Positive
HateXplain	
We cannot continue calling ourselves feminists if the rights of all womxn arent addressed yes to a sexual offences public list but will a trans lesbian bisexual and queer womxn be able to enter their information on the reporting sheet gender forum	Neutral
First wave feminists were largely thoughtful women advocating for voting rights legal financial rights access to opportunity subsequent feminist movements have been ridiculou by equating feminism with abortion sexuality etc burning ones bra never accomplished anything except sagging	Hate speech
VAST	
The era when there were no public schools was not a good socio-economic time in the life of our nation. Anything which weakens public schools and their funding will result in the most vulnerable youth of America being left out of the chance to get an education. [SEP] home schoolers	Pro
Our school district is terrrible, and the results even worse. Last year a child was T-Bagged and beaten because he was white by a group of black football players. No one was punished...b/c they were afraid of the moron parents. We send all 3 kids to private school at a considerable expense to ourselves yet the money we pay in taxes goes to the school we had to pull our kids from. Is that fair? [SEP] home schoolers	Con

Table 4: Examples for the IMDb, SST, HateXplain, and VAST datasets. The right column shows the labels. Human-annotated rationales are underlined.

G.2 DeBERTa

See Tables 21-24 for the OOD performance for DeBERTa models.

H Spearman rank correlation between metrics

In Figure 4 we show the Spearman rank correlation between metrics for BERT across datasets.

	train	dev	test
pos	702	89	93
neg	770	93	96
total	1472	182	189

(a) IMDb

	train	dev	test
pos	3440	423	875
neg	3101	411	848
total	6541	834	1723

(b) SST

	train	dev	test
abusive	6429	1640	1105
hate	7489	1880	1314
total	14102	3627	2418

(c) HateXplain

	train	dev	test
pro	249	11	10
con	270	11	11
total	519	22	21

(d) VAST

Table 5: Number of instances per dataset after preprocessing.

	IMDb	SST	HX	VAST
tokens	489	22	27	128
rationales	33	10	11	32
rationale ratio	0.06	0.46	0.43	0.25

(a) Average number of tokens, average number of rationale tokens, and token-rationale ratio across datasets.

	IMDb	SST	HX	VAST
$\mathcal{L}_{KLD}$	8.4×10^{-2}	9.8×10^{-2}	8.5×10^{-2}	1.2×10^{-1}
$\mathcal{L}_{AM}$	9.9×10^{-1}	5.4×10^{-1}	9.7×10^{-1}	9.9×10^{-1}
$\mathcal{L}_{MSE}$	8.0×10^{-7}	2.0×10^{-5}	7.0×10^{-4}	5.0×10^{-6}
$\mathcal{L}_{MAE}$	7.2×10^{-4}	7.5×10^{-4}	2.6×10^{-3}	8.0×10^{-4}
$\mathcal{L}_{OL}$	1.0×10^{-4}	5.9×10^{-2}	2.0×10^{-4}	7.0×10^{-5}

(b) $\mathcal{L}_r$ calculated on the test set using different loss functions on an untrained model.

Table 6: Rationale-token ratio and $\mathcal{L}_r$. HX refers to the HateXplain dataset. $\mathcal{L}_{MSE}$, $\mathcal{L}_{MAE}$, and $\mathcal{L}_{OL}$, are sensitive to token length.

ID - IMDb - BERT

DS		100			200			500			1472		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}
None		.57 ± .04	.48	.33	.48 ± .09	.48	.48	.74 ± .16	.53	.53	.92 ± .03	.64	.61
$\mathcal{L}_{EAR}$		.58 ± .03	.48	.49	.49 ± .1	.48	.47	.73 ± .16	.51	.53	.93 ± .01	.63	.60
$\mathcal{L}_{AM}$	h	.61 ± .01	.48	.49	.46 ± .09	.48	.49	.91 ± .01	.59	.61	.94 ± .01	.64	.62
	r	.59 ± .0	.49	.53	.54 ± .0	.47	.50	.88 ± .05	.61	.59	.88 ± .01	.62	.61
$\mathcal{L}_{KLD}$	h	.62 ± .07	.54	.51	.75 ± .05	.58	.54	.92 ± .02	.62	.62	.91 ± .02	.70	.60
	r	.51 ± .01	.48	.55	.74 ± .05	.51	.55	.92 ± .04	.61	.60	.94 ± .04	.55	.61
$\mathcal{L}_{MAE}$	h	.67 ± .02	.48	.49	.73 ± .04	.53	.52	.81 ± .07	.62	.53	.95 ± .01	.63	.58
	r	.63 ± .01	.50	.54	.63 ± .13	.50	.52	.92 ± .01	.56	.55	.94 ± .03	.65	.60
$\mathcal{L}_{MSE}$	h	.71 ± .06	.68	.51	.58 ± .09	.53	.49	.92 ± .02	.62	.61	.92 ± .03	.63	.61
	r	.41 ± .14	.50	.50	.61 ± .03	.47	.51	.8 ± .07	.61	.57	.89 ± .02	.62	.62
$\mathcal{L}_{OL}$	h	.51 ± .03	.51	.50	.78 ± .03	.60	.55	.89 ± .03	.62	.59	.92 ± .01	.65	.62
	r	.6 ± .01	.50	.51	.67 ± .2	.54	.55	.85 ± .02	.58	.58	.9 ± .0	.61	.62

Table 7: In-domain performance for IMDb with BERT. $\boldsymbol{a}^h$ refers to the attention pattern type as human-annotated rationale tokens (h) or randomly selected tokens (r). $F1$ denotes the F1-score on the ID test set. $\pm$ denotes the standard deviation. A_{AUC} denotes the AUC for extracted attention patterns. S_{AUC} denotes the AUC for SHAP explanations.

ID - SST - BERT

DS		100			200			500			6541		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}
None		.55 ± .14	.36	.52	.37 ± .07	.36	.54	.85 ± .01	.42	.60	.89 ± .0	.49	.64
$\mathcal{L}_{EAR}$		.48 ± .21	.36	.50	.37 ± .07	.36	.54	.86 ± .01	.44	.61	.9 ± .0	.46	.60
$\mathcal{L}_{AM}$	h	.72 ± .0	.80	.50	.86 ± .0	.79	.61	.87 ± .0	.79	.62	.91 ± .0	.74	.62
	r	.73 ± .0	.53	.50	.89 ± .0	.53	.60	.88 ± .0	.52	.62	.89 ± .0	.53	.62
$\mathcal{L}_{KLD}$	h	.45 ± .0	.83	.47	.46 ± .0	.36	.48	.89 ± .0	.80	.62	.92 ± .0	.83	.61
	r	.44 ± .1	.59	.48	.47 ± .0	.59	.48	.86 ± .02	.59	.60	.9 ± .02	.59	.61
$\mathcal{L}_{MAE}$	h	.62 ± .0	.67	.47	.87 ± .0	.59	.62	.87 ± .0	.54	.58	.92 ± .0	.83	.61
	r	.62 ± .0	.54	.48	.84 ± .01	.54	.61	.87 ± .0	.54	.62	.9 ± .02	.54	.61
$\mathcal{L}_{MSE}$	h	.57 ± .0	.82	.49	.87 ± .0	.75	.59	.87 ± .0	.82	.63	.91 ± .0	.77	.60
	r	.43 ± .09	.53	.51	.86 ± .01	.53	.58	.86 ± .02	.53	.61	.9 ± .01	.53	.61
$\mathcal{L}_{OL}$	h	.62 ± .0	.75	.47	.86 ± .0	.74	.60	.88 ± .0	.72	.61	.91 ± .0	.77	.61
	r	.7 ± .0	.53	.49	.86 ± .0	.53	.61	.86 ± .0	.53	.61	.91 ± .01	.53	.62

Table 8: In-domain performance for SST with BERT. $\boldsymbol{a}^h$ refers to the attention pattern type as human-annotated rationale tokens (h) or randomly selected tokens (r). $F1$ denotes F1-score on the ID test set. $\pm$ denotes the standard deviation. A_{AUC} denotes the AUC for extracted attention patterns. S_{AUC} denotes the AUC for SHAP explanations.

ID - HateXplain - BERT

DS		100			200			500			14102		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}
None		.55 ± .14	.36	.52	.37 ± .07	.36	.54	.85 ± .01	.42	.60	.89 ± .0	.49	.64
$\mathcal{L}_{EAR}$		.48 ± .21	.36	.50	.37 ± .07	.36	.54	.86 ± .01	.44	.61	.9 ± .0	.46	.60
$\mathcal{L}_{AM}$	h	.72 ± .0	.80	.50	.86 ± .0	.79	.61	.87 ± .0	.79	.62	.91 ± .0	.74	.62
	r	.73 ± .0	.53	.50	.89 ± .0	.53	.60	.88 ± .0	.52	.62	.89 ± .0	.53	.62
$\mathcal{L}_{KLD}$	h	.45 ± .0	.83	.47	.46 ± .0	.36	.48	.89 ± .0	.80	.62	.92 ± .0	.83	.61
	r	.44 ± .1	.59	.48	.47 ± .0	.59	.48	.86 ± .02	.59	.60	.9 ± .02	.59	.61
$\mathcal{L}_{MAE}$	h	.62 ± .0	.67	.47	.87 ± .0	.59	.62	.87 ± .0	.54	.58	.92 ± .0	.83	.61
	r	.62 ± .0	.54	.48	.84 ± .01	.54	.61	.87 ± .0	.54	.62	.9 ± .02	.54	.61
$\mathcal{L}_{MSE}$	h	.57 ± .0	.82	.49	.87 ± .0	.75	.59	.87 ± .0	.82	.63	.91 ± .0	.77	.60
	r	.43 ± .09	.53	.51	.86 ± .01	.53	.58	.86 ± .02	.53	.61	.9 ± .01	.53	.61
$\mathcal{L}_{OL}$	h	.62 ± .0	.75	.47	.86 ± .0	.74	.60	.88 ± .0	.72	.61	.91 ± .0	.77	.61
	r	.7 ± .0	.53	.49	.86 ± .0	.53	.61	.86 ± .0	.53	.61	.91 ± .01	.53	.62

Table 9: In-domain performance for HateXplain with BERT. $\boldsymbol{a}^h$ refers to the attention pattern type as human-annotated rationale tokens (h) or randomly selected tokens (r). $F1$ denotes F1-score on the ID test set. $\pm$ denotes the standard deviation.A_{AUC} denotes the AUC for extracted attention patterns. S_{AUC} denotes the AUC for SHAP explanations.

ID - VAST - BERT

DS		100			200			519		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}
None		.39 ± .05	.54	.51	.47 ± .0	.56	.51	.42 ± .0	.55	.55
$\mathcal{L}_{EAR}$		.39 ± .05	.54	.51	.38 ± .0	.56	.51	.42 ± .0	.55	.54
$\mathcal{L}_{AM}$	h	.45 ± .0	.55	.49	.43 ± .0	.54	.53	.42 ± .0	.54	.51
	r	.45 ± .0	.55	.50	.49 ± .05	.54	.54	.42 ± .0	.54	.51
$\mathcal{L}_{KLD}$	h	.42 ± .0	.54	.51	.47 ± .0	.55	.57	.38 ± .0	.54	.51
	r	.42 ± .0	.54	.50	.44 ± .03	.55	.55	.42 ± .0	.54	.51
$\mathcal{L}_{MAE}$	h	.46 ± .0	.64	.49	.46 ± .0	.56	.57	.39 ± .0	.56	.55
	r	.44 ± .02	.54	.50	.48 ± .03	.54	.54	.44 ± .05	.55	.53
$\mathcal{L}_{MSE}$	h	.53 ± .0	.64	.48	.46 ± .0	.55	.52	.5 ± .0	.55	.51
	r	.43 ± .02	.55	.50	.46 ± .0	.55	.52	.5 ± .0	.54	.51
$\mathcal{L}_{OL}$	h	.57 ± .0	.65	.50	.3 ± .0	.59	.52	.57 ± .0	.64	.53
	r	.56 ± .13	.53	.53	.48 ± .16	.55	.52	.53 ± .18	.55	.56

Table 10: In-domain performance for VAST with BERT. $\boldsymbol{a}^h$ refers to the attention pattern type as human-annotated rationale tokens (h) or randomly selected tokens (r). $F1$ denotes F1-score on the ID test set. $\pm$ denotes the standard deviation. A_{AUC} denotes the AUC for extracted attention patterns. S_{AUC} denotes the AUC for SHAP explanations.

ID - IMDb - DeBERTa

DS		100			200			500			1472		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}
None		.91 ± .01	.58	.54	.95 ± .01	.55	.58	.94 ± .01	.55	.61	.96 ± .0	.65	.61
$\mathcal{L}_{EAR}$		.92 ± .01	.53	.54	.95 ± .01	.55	.58	.93 ± .03	.58	.61	.95 ± .01	.57	.61
$\mathcal{L}_{AM}$	h	.91 ± .01	.56	.54	.92 ± .03	.63	.47	.95 ± .01	.55	.60	.96 ± .01	.59	.62
	r	.91 ± .01	.50	.54	.94 ± .0	.50	.47	.95 ± .02	.50	.61	.96 ± .01	.50	.61
$\mathcal{L}_{KLD}$	h	.9 ± .01	.61	.56	.94 ± .03	.58	.58	.95 ± .02	.61	.56	.93 ± .0	.61	.59
	r	.91 ± .01	.50	.56	.93 ± .01	.49	.58	.94 ± .02	.49	.57	.97 ± .0	.51	.58
$\mathcal{L}_{MAE}$	h	.87 ± .02	.54	.56	.94 ± .0	.61	.57	.95 ± .01	.62	.61	.95 ± .01	.58	.61
	r	.89 ± .04	.50	.56	.95 ± .01	.50	.57	.92 ± .02	.50	.61	.95 ± .01	.50	.61
$\mathcal{L}_{MSE}$	h	.93 ± .01	.62	.52	.95 ± .0	.58	.59	.95 ± .01	.63	.55	.95 ± .01	.61	.59
	r	.9 ± .02	.50	.52	.94 ± .01	.50	.59	.95 ± .02	.50	.58	.96 ± .0	.50	.61
$\mathcal{L}_{OL}$	h	.34 ± .0	.56	.54	.4 ± .11	.53	.58	.94 ± .04	.57	.61	.96 ± .0	.55	.59
	r	.35 ± .01	.50	.57	.44 ± .17	.50	.58	.9 ± .03	.50	.60	.96 ± .0	.50	.58

Table 11: In-domain performance for IMDb with DeBERTa. $\boldsymbol{a}^h$ refers to the attention pattern type as human-annotated rationale tokens (h) or randomly selected tokens (r). $F1$ denotes the F1-score on the ID test set. $\pm$ denotes the standard deviation. A_{AUC} denotes the AUC for extracted attention patterns. S_{AUC} denotes the AUC for SHAP explanations.

ID - SST - DeBERTa

DS		100			200			500			6541		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}
None		.34 ± .0	.59	.52	.58 ± .06	.54	.56	.92 ± .01	.55	.61	.94 ± .01	.53	.57
$\mathcal{L}_{EAR}$		.5 ± .28	.52	.55	.65 ± .09	.52	.56	.91 ± .0	.48	.61	.95 ± .01	.49	.57
$\mathcal{L}_{AM}$	h	.68 ± .3	.60	.54	.91 ± .02	.64	.54	.86 ± .01	.72	.55	.95 ± .01	.56	.60
	r	.65 ± .27	.50	.54	.92 ± .01	.50	.54	.92 ± .01	.50	.55	.95 ± .01	.50	.60
$\mathcal{L}_{KLD}$	h	.34 ± .01	.63	.54	.88 ± .02	.71	.54	.92 ± .01	.72	.59	.95 ± .0	.74	.56
	r	.37 ± .06	.53	.54	.92 ± .0	.58	.54	.93 ± .0	.59	.59	.95 ± .01	.60	.56
$\mathcal{L}_{MAE}$	h	.33 ± .0	.55	.52	.88 ± .01	.72	.54	.92 ± .0	.64	.58	.95 ± .0	.73	.58
	r	.33 ± .0	.51	.52	.9 ± .02	.51	.54	.92 ± .0	.50	.58	.95 ± .0	.52	.58
$\mathcal{L}_{MSE}$	h	.46 ± .1	.71	.54	.91 ± .01	.71	.54	.92 ± .0	.68	.56	.96 ± .0	.72	.57
	r	.34 ± .0	.50	.54	.89 ± .02	.50	.54	.92 ± .0	.51	.56	.95 ± .0	.51	.57
$\mathcal{L}_{OL}$	h	.33 ± .0	.55	.53	.33 ± .0	.56	.53	.88 ± .04	.62	.58	.94 ± .01	.63	.59
	r	.33 ± .0	.50	.53	.33 ± .0	.50	.53	.93 ± .01	.50	.58	.95 ± .01	.50	.59

Table 12: In-domain performance for SST with DeBERTa. $\boldsymbol{a}^h$ refers to the attention pattern type as human-annotated rationale tokens (h) or randomly selected tokens (r). $F1$ denotes F1-score on the ID test set. $\pm$ denotes the standard deviation. A_{AUC} denotes the AUC for extracted attention patterns. S_{AUC} denotes the AUC for SHAP explanations.

ID - HateXplain - DeBERTa

DS		100			200			500			14102		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}
None		.35 ± .0	.51	.57	.35 ± .0	.52	.58	.39 ± .03	.59	.58	.68 ± .01	.58	.57
$\mathcal{L}_{EAR}$		.35 ± .0	.53	.57	.35 ± .0	.54	.58	.47 ± .1	.55	.58	.74 ± .0	.49	.57
$\mathcal{L}_{AM}$	h	.46 ± .18	.59	.57	.35 ± .32	.51	.58	.7 ± .01	.79	.64	.73 ± .01	.80	.75
	r	.46 ± .18	.50	.57	.35 ± .32	.50	.58	.45 ± .17	.50	.57	.73 ± .0	.50	.74
$\mathcal{L}_{KLD}$	h	.34 ± .33	.55	.57	.36 ± .31	.51	.62	.71 ± .0	.80	.72	.72 ± .0	.82	.68
	r	.34 ± .33	.51	.57	.36 ± .31	.51	.64	.45 ± .16	.56	.70	.71 ± .01	.56	.65
$\mathcal{L}_{MAE}$	h	.46 ± .18	.65	.57	.45 ± .17	.64	.59	.71 ± .01	.78	.69	.73 ± .01	.81	.73
	r	.46 ± .18	.50	.57	.46 ± .19	.50	.63	.57 ± .19	.50	.70	.73 ± .01	.51	.72
$\mathcal{L}_{MSE}$	h	.36 ± .3	.51	.57	.46 ± .14	.79	.66	.69 ± .02	.79	.71	.72 ± .01	.82	.72
	r	.36 ± .3	.50	.57	.35 ± .0	.50	.68	.64 ± .05	.50	.71	.71 ± .01	.50	.70
$\mathcal{L}_{OL}$	h	.35 ± .32	.51	.58	.35 ± .0	.66	.65	.43 ± .13	.78	.66	.73 ± .01	.66	.71
	r	.35 ± .32	.50	.58	.35 ± .0	.50	.68	.44 ± .14	.50	.70	.74 ± .01	.50	.69

Table 13: In-domain performance for HateXplain with DeBERTa. $\boldsymbol{a}^h$ refers to the attention pattern type as human-annotated rationale tokens (h) or randomly selected tokens (r). $F1$ denotes F1-score on the ID test set. ± denotes the standard deviation.A_{AUC} denotes the AUC for extracted attention patterns. S_{AUC} denotes the AUC for SHAP explanations.

ID - VAST - DeBERTa

DS		100			200			519		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}	$F1$	A_{AUC}	S_{AUC}
None		.44 ± .16	.58	.48	.61 ± .04	.58	.50	.74 ± .03	.58	.44
$\mathcal{L}_{EAR}$		.47 ± .13	.54	.48	.64 ± .07	.55	.50	.69 ± .08	.55	.45
$\mathcal{L}_{AM}$	h	.58 ± .07	.65	.49	.56 ± .13	.56	.56	.69 ± .11	.65	.52
	r	.58 ± .21	.51	.49	.56 ± .13	.50	.56	.59 ± .12	.51	.52
$\mathcal{L}_{KLD}$	h	.68 ± .11	.60	.45	.61 ± .04	.63	.53	.71 ± .09	.65	.45
	r	.67 ± .08	.49	.45	.48 ± .11	.49	.53	.6 ± .24	.50	.45
$\mathcal{L}_{MAE}$	h	.63 ± .1	.60	.51	.64 ± .04	.64	.53	.75 ± .03	.58	.49
	r	.69 ± .06	.50	.51	.66 ± .06	.50	.53	.61 ± .25	.50	.49
$\mathcal{L}_{MSE}$	h	.55 ± .09	.64	.46	.56 ± .01	.63	.43	.54 ± .11	.56	.44
	r	.64 ± .09	.50	.46	.64 ± .03	.49	.43	.55 ± .2	.50	.44
$\mathcal{L}_{OL}$	h	.33 ± .01	.54	.51	.37 ± .05	.54	.51	.32 ± .0	.56	.51
	r	.33 ± .01	.50	.51	.32 ± .01	.51	.51	.48 ± .17	.51	.51

Table 14: In-domain performance for VAST with DeBERTa. $\boldsymbol{a}^h$ refers to the attention pattern type as human-annotated rationale tokens (h) or randomly selected tokens (r). $F1$ denotes F1-score on the ID test set. ± denotes the standard deviation. A_{AUC} denotes the AUC for extracted attention patterns. S_{AUC} denotes the AUC for SHAP explanations.

Spurious correlations - IMDb - BERT

$SC_{\%}$		25%					50%					75%				
$\mathcal{L}_r$	$\boldsymbol{a}^h$	Sp_{F1}	F_{F1}	N_{F1}	A_{AUC}	S_{AUC}	Sp_{F1}	F_{F1}	N_{F1}	A_{AUC}	S_{AUC}	Sp_{F1}	F_{F1}	N_{F1}	A_{AUC}	S_{AUC}
None		1.0 ± .0	.07 ± .04	.89 ± .0	.61 ± .01	.59 ± .0	1.0 ± .0	.0 ± .0	.83 ± .0	.58 ± .03	.55 ± .01	1.0 ± .0	.0 ± .0	.55 ± .0	.49 ± .02	.54 ± .02
$\mathcal{L}_{EAR}$		1.0 ± .0	.05 ± .08	.78 ± .0	.54 ± .06	.55 ± .02	1.0 ± .0	.0 ± .0	.78 ± .0	.55 ± .07	.55 ± .02	1.0 ± .0	.0 ± .0	.57 ± .0	.49 ± .03	.52 ± .03
$\mathcal{L}_{AM}$	h	.99 ± .0	.23 ± .0	.88 ± .0	.66 ± .04	.57 ± .04	1.0 ± .0	.0 ± .0	.79 ± .0	.56 ± .04	.56 ± .04	1.0 ± .0	.0 ± .0	.64 ± .0	.54 ± .04	.53 ± .04
	r	1.0 ± .0	.11 ± .08	.91 ± .0	.61 ± .02	.57 ± .01	1.0 ± .0	.0 ± .0	.71 ± .0	.54 ± .03	.54 ± .02	1.0 ± .0	.0 ± .0	.59 ± .0	.48 ± .02	.49 ± .02
$\mathcal{L}_{KLD}$	h	1.0 ± .0	.11 ± .0	.91 ± .0	.67 ± .0	.59 ± .01	1.0 ± .0	.06 ± .0	.88 ± .0	.66 ± .0	.58 ± .01	1.0 ± .0	.0 ± .0	.82 ± .0	.69 ± .0	.54 ± .01
	r	1.0 ± .0	.08 ± .07	.89 ± .0	.59 ± .03	.59 ± .01	1.0 ± .0	.0 ± .0	.73 ± .0	.5 ± .01	.53 ± .01	1.0 ± .0	.0 ± .0	.62 ± .0	.48 ± .01	.51 ± .01
$\mathcal{L}_{MAE}$	h	1.0 ± .0	.1 ± .0	.9 ± .0	.6 ± .03	.58 ± .03	1.0 ± .0	.05 ± .0	.87 ± .0	.65 ± .03	.56 ± .03	1.0 ± .0	.0 ± .0	.78 ± .0	.61 ± .03	.54 ± .03
	r	1.0 ± .0	.08 ± .07	.9 ± .0	.57 ± .02	.56 ± .01	1.0 ± .0	.0 ± .0	.86 ± .0	.56 ± .03	.55 ± .03	1.0 ± .0	.0 ± .0	.43 ± .0	.47 ± .01	.51 ± .02
$\mathcal{L}_{MSE}$	h	1.0 ± .0	.27 ± .0	.9 ± .0	.69 ± .0	.58 ± .02	1.0 ± .0	.07 ± .0	.86 ± .0	.7 ± .0	.57 ± .02	1.0 ± .0	.01 ± .0	.83 ± .0	.69 ± .0	.55 ± .02
	r	1.0 ± .0	.09 ± .02	.86 ± .0	.54 ± .03	.56 ± .01	1.0 ± .0	.0 ± .0	.8 ± .0	.52 ± .02	.53 ± .01	1.0 ± .0	.0 ± .0	.57 ± .0	.47 ± .01	.52 ± .01
$\mathcal{L}_{OL}$	h	1.0 ± .0	.07 ± .0	.87 ± .0	.63 ± .01	.57 ± .03	1.0 ± .0	.01 ± .0	.75 ± .0	.61 ± .01	.56 ± .03	1.0 ± .0	.0 ± .0	.76 ± .0	.63 ± .01	.55 ± .03
	r	1.0 ± .0	.07 ± .06	.84 ± .0	.61 ± .01	.58 ± .01	1.0 ± .0	.02 ± .03	.85 ± .0	.57 ± .02	.57 ± .03	1.0 ± .0	.0 ± .0	.77 ± .0	.58 ± .01	.55 ± .03

Table 15: Averaged results for BERT trained on 500 examples from the IMDb dataset with spurious correlations. $SC_{\%}$ denotes the percentage of synthetic spurious correlations in the training set. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r). $\pm$ denotes the standard deviation. Sp_{F1} refers to the F1-score on the dataset with spurious correlations, F_{F1} to the dataset with flipped correlations, and N_{F1} to the dataset without spurious correlations.

Spurious correlations - IMDb - DeBERTa

$SC_{\%}$			25%				50%				75%			
$\mathcal{L}_r$	$\boldsymbol{a}^h$	Sp_{F1}	F_{F1}	N_{F1}	A_{AUC}	S_{AUC}	Sp_{F1}	F_{F1}	N_{F1}	A_{AUC}	S_{AUC}	Sp_{F1}	F_{F1}	N_{F1}	A_{AUC}	S_{AUC}
None		.95 ± .05	.51 ± .26	.89 ± .05	.63 ± .01	.57 ± .03	1.0 ± .0	.16 ± .11	.9 ± .0	.58 ± .03	.58 ± .0	1.0 ± .0	.05 ± .09	.88 ± .0	.58 ± .02	.56 ± .03
$\mathcal{L}_{EAR}$		.95 ± .05	.51 ± .26	.89 ± .05	.63 ± .01	.57 ± .03	1.0 ± .0	.16 ± .11	.9 ± .0	.58 ± .03	.58 ± .0	1.0 ± .0	.05 ± .09	.88 ± .0	.58 ± .02	.56 ± .03
$\mathcal{L}_{AM}$	h	.99 ± .0	.47 ± .11	.94 ± .0	.56 ± .0	.6 ± .01	.99 ± .0	.34 ± .11	.94 ± .0	.53 ± .0	.58 ± .01	1.0 ± .0	.08 ± .11	.93 ± .0	.5 ± .0	.57 ± .01
	r	.99 ± .01	.43 ± .29	.94 ± .01	.49 ± .0	.58 ± .04	.99 ± .01	.31 ± .16	.93 ± .01	.49 ± .0	.58 ± .02	1.0 ± .0	.08 ± .12	.93 ± .0	.49 ± .0	.57 ± .01
$\mathcal{L}_{KLD}$	h	.97 ± .0	.65 ± .11	.94 ± .0	.59 ± .01	.59 ± .02	1.0 ± .0	.17 ± .11	.93 ± .0	.57 ± .01	.55 ± .02	1.0 ± .0	.06 ± .11	.93 ± .0	.53 ± .01	.56 ± .02
	r	.99 ± .01	.56 ± .21	.94 ± .01	.48 ± .0	.59 ± .02	1.0 ± .0	.13 ± .05	.94 ± .0	.48 ± .0	.59 ± .01	1.0 ± .0	.06 ± .11	.93 ± .0	.49 ± .0	.57 ± .01
$\mathcal{L}_{MAE}$	h	.98 ± .01	.73 ± .24	.89 ± .01	.59 ± .0	.58 ± .02	1.0 ± .01	.3 ± .24	.95 ± .01	.55 ± .0	.6 ± .02	.99 ± .01	.37 ± .24	.94 ± .01	.61 ± .0	.57 ± .02
	r	.98 ± .03	.52 ± .34	.94 ± .03	.5 ± .0	.58 ± .02	1.0 ± .01	.21 ± .1	.94 ± .01	.49 ± .0	.59 ± .05	1.0 ± .0	.2 ± .03	.93 ± .0	.5 ± .01	.55 ± .01
$\mathcal{L}_{MSE}$	h	.98 ± .0	.55 ± .08	.94 ± .0	.62 ± .0	.59 ± .03	1.0 ± .0	.17 ± .08	.85 ± .0	.6 ± .0	.56 ± .03	1.0 ± .0	.06 ± .08	.86 ± .0	.57 ± .0	.56 ± .03
	r	.98 ± .0	.55 ± .23	.94 ± .0	.62 ± .01	.59 ± .01	1.0 ± .0	.17 ± .11	.85 ± .0	.6 ± .02	.56 ± .02	1.0 ± .0	.06 ± .08	.86 ± .0	.57 ± .0	.56 ± .03
$\mathcal{L}_{OL}$	h	.97 ± .19	.71 ± .28	.94 ± .19	.58 ± .03	.56 ± .04	.99 ± .19	.45 ± .28	.92 ± .19	.55 ± .03	.55 ± .04	.85 ± .19	.45 ± .28	.69 ± .19	.54 ± .03	.55 ± .04
	r	.97 ± .01	.74 ± .15	.93 ± .01	.5 ± .0	.57 ± .02	.99 ± .01	.45 ± .28	.9 ± .01	.5 ± .0	.56 ± .04	.77 ± .38	.31 ± .27	.73 ± .38	.5 ± .0	.54 ± .05

Table 16: Averaged results for DeBERTa trained on 500 examples from the IMDb dataset with spurious correlations. $SC_{\%}$ denotes the percentage of synthetic spurious correlations in the training set. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r). $\pm$ denotes the standard deviation. Sp_{F1} refers to the F1-score on the dataset with spurious correlations, F_{F1} to the dataset with flipped correlations, and N_{F1} to the dataset without spurious correlations.

OOD - IMDb - BERT

$\mathcal{L}_r$	DS / $\boldsymbol{a}^h$	100				200				500				1472			
		Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU
None		.42 ± .11	.48 ± .11	.46 ± .09	.42 ± .12	.41 ± .1	.51 ± .12	.44 ± .08	.4 ± .1	.48 ± .17	.61 ± .24	.53 ± .17	.45 ± .18	.68 ± .02	.87 ± .02	.79 ± .04	.68 ± .02
$\mathcal{L}_{EAR}$		.41 ± .1	.47 ± .1	.42 ± .04	.4 ± .08	.41 ± .11	.5 ± .11	.43 ± .07	.4 ± .1	.49 ± .19	.62 ± .26	.53 ± .18	.45 ± .18	.69 ± .02	.88 ± .02	.78 ± .03	.69 ± .0
$\mathcal{L}_{AM}$	h	.5 ± .07	.49 ± .09	.49 ± .09	.48 ± .07	.41 ± .11	.51 ± .13	.43 ± .1	.4 ± .1	.7 ± .0	.88 ± .01	.78 ± .02	.67 ± .0	.69 ± .01	.88 ± .02	.8 ± .02	.69 ± .01
	r	.43 ± .0	.54 ± .0	.61 ± .0	.45 ± .0	.51 ± .0	.52 ± .0	.5 ± .0	.48 ± .0	.67 ± .02	.88 ± .02	.77 ± .03	.67 ± .0	.66 ± .08	.82 ± .12	.78 ± .06	.68 ± .05
$\mathcal{L}_{KLD}$	h	.53 ± .08	.54 ± .1	.51 ± .1	.52 ± .08	.53 ± .11	.71 ± .12	.68 ± .07	.58 ± .1	.7 ± .01	.88 ± .01	.8 ± .0	.67 ± .02	.6 ± .1	.87 ± .03	.8 ± .02	.65 ± .05
	r	.39 ± .0	.48 ± .01	.6 ± .01	.41 ± .01	.56 ± .09	.65 ± .19	.62 ± .07	.61 ± .07	.67 ± .01	.88 ± .01	.78 ± .02	.68 ± .01	.7 ± .01	.89 ± .01	.8 ± .04	.7 ± .0
$\mathcal{L}_{MAE}$	h	.52 ± .09	.59 ± .1	.59 ± .05	.52 ± .12	.53 ± .13	.71 ± .16	.67 ± .07	.55 ± .11	.48 ± .12	.66 ± .23	.64 ± .11	.45 ± .1	.63 ± .11	.88 ± .03	.82 ± .02	.65 ± .07
	r	.5 ± .01	.63 ± .02	.61 ± .01	.53 ± .01	.5 ± .11	.54 ± .22	.48 ± .14	.53 ± .12	.61 ± .01	.78 ± .02	.72 ± .03	.53 ± .01	.71 ± .01	.9 ± .02	.81 ± .03	.69 ± .01
$\mathcal{L}_{MSE}$	h	.63 ± .02	.73 ± .07	.68 ± .06	.6 ± .08	.45 ± .09	.54 ± .12	.45 ± .1	.43 ± .1	.69 ± .02	.88 ± .0	.79 ± .02	.68 ± .01	.7 ± .01	.88 ± .02	.79 ± .04	.7 ± .01
	r	.45 ± .04	.4 ± .02	.42 ± .14	.46 ± .05	.51 ± .03	.58 ± .01	.54 ± .03	.45 ± .03	.7 ± .0	.84 ± .02	.69 ± .04	.69 ± .0	.7 ± .01	.86 ± .02	.78 ± .02	.69 ± .01
$\mathcal{L}_{OL}$	h	.4 ± .1	.45 ± .11	.4 ± .01	.39 ± .1	.7 ± .02	.84 ± .03	.68 ± .06	.66 ± .02	.72 ± .01	.9 ± .01	.78 ± .01	.68 ± .01	.7 ± .03	.89 ± .02	.8 ± .02	.7 ± .01
	r	.55 ± .0	.55 ± .0	.45 ± .0	.56 ± .0	.6 ± .2	.71 ± .31	.69 ± .1	.55 ± .16	.71 ± .01	.84 ± .02	.75 ± .02	.68 ± .01	.68 ± .04	.86 ± .01	.78 ± .03	.68 ± .02

Table 17: Out-of-domain performance for models trained on IMDb with BERT. We report macro average F1-scores for Dynasent 1 (Ds1), Dynasent 2 (Ds2), Yelp (Y), and SFU review (SFU). DS denotes the number of instances in the IMDb training set. ± denotes the standard deviation. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r).

OOD - SST - BERT

DS		100				200				500				1472			
$\mathcal{L}_r$	$\boldsymbol{a}^h$	Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU
None		.49 ± .08	.49 ± .14	.46 ± .11	.47 ± .1	.37 ± .05	.34 ± .01	.35 ± .0	.36 ± .04	.72 ± .01	.87 ± .02	.73 ± .05	.72 ± .01	.68 ± .0	.89 ± .0	.78 ± .0	.7 ± .0
$\mathcal{L}_{EAR}$		.44 ± .14	.43 ± .19	.41 ± .13	.43 ± .14	.37 ± .05	.34 ± .01	.35 ± .0	.36 ± .04	.72 ± .01	.88 ± .0	.73 ± .01	.7 ± .01	.64 ± .0	.88 ± .0	.69 ± .0	.68 ± .0
$\mathcal{L}_{AM}$	h	.66 ± .0	.74 ± .0	.64 ± .0	.58 ± .0	.64 ± .0	.82 ± .0	.67 ± .0	.65 ± .0	.69 ± .0	.87 ± .0	.68 ± .0	.68 ± .0	.69 ± .0	.9 ± .0	.79 ± .0	.7 ± .0
	r	.64 ± .0	.72 ± .0	.62 ± .0	.57 ± .0	.7 ± .0	.85 ± .0	.62 ± .0	.69 ± .0	.7 ± .0	.89 ± .0	.73 ± .0	.7 ± .0	.7 ± .0	.88 ± .0	.77 ± .0	.71 ± .0
$\mathcal{L}_{KLD}$	h	.47 ± .0	.42 ± .0	.44 ± .0	.4 ± .0	.43 ± .0	.34 ± .0	.35 ± .0	.41 ± .0	.68 ± .0	.86 ± .0	.66 ± .0	.71 ± .0	.7 ± .0	.92 ± .0	.78 ± .0	.72 ± .0
	r	.39 ± .07	.34 ± .03	.36 ± .01	.37 ± .05	.43 ± .0	.34 ± .0	.35 ± .0	.41 ± .0	.71 ± .01	.88 ± .01	.75 ± .01	.7 ± .01	.71 ± .0	.9 ± .0	.78 ± .03	.72 ± .01
$\mathcal{L}_{MAE}$	h	.56 ± .0	.52 ± .0	.47 ± .0	.46 ± .0	.67 ± .01	.86 ± .01	.7 ± .01	.68 ± .0	.71 ± .0	.89 ± .01	.73 ± .01	.7 ± .01	.7 ± .01	.9 ± .01	.79 ± .04	.73 ± .01
	r	.56 ± .0	.52 ± .0	.48 ± .0	.46 ± .0	.7 ± .01	.85 ± .01	.7 ± .01	.7 ± .0	.73 ± .0	.89 ± .01	.77 ± .01	.7 ± .01	.68 ± .01	.89 ± .01	.75 ± .01	.71 ± .01
$\mathcal{L}_{MSE}$	h	.53 ± .0	.57 ± .0	.53 ± .0	.45 ± .0	.71 ± .0	.9 ± .0	.76 ± .0	.69 ± .0	.69 ± .0	.86 ± .0	.68 ± .0	.71 ± .0	.69 ± .0	.9 ± .0	.8 ± .0	.73 ± .0
	r	.38 ± .04	.35 ± .05	.35 ± .01	.36 ± .03	.71 ± .01	.87 ± .01	.72 ± .03	.71 ± .01	.73 ± .0	.88 ± .01	.74 ± .02	.7 ± .01	.7 ± .01	.91 ± .01	.8 ± .01	.71 ± .01
$\mathcal{L}_{OL}$	h	.55 ± .0	.47 ± .0	.41 ± .0	.47 ± .0	.68 ± .0	.87 ± .0	.73 ± .0	.68 ± .0	.67 ± .0	.82 ± .0	.58 ± .0	.69 ± .0	.7 ± .0	.91 ± .0	.81 ± .0	.74 ± .0
	r	.61 ± .0	.63 ± .0	.59 ± .0	.51 ± .0	.72 ± .0	.87 ± .0	.65 ± .0	.7 ± .0	.71 ± .0	.88 ± .0	.73 ± .0	.7 ± .0	.68 ± .02	.9 ± .01	.75 ± .03	.7 ± .01

Table 18: Out-of-domain performance for models trained on SST with BERT. We report macro average F1-scores for Dynasent 1 (Ds1), Dynasent 2 (Ds2), Yelp (Y), and SFU review (SFU). DS denotes the number of instances in the SST training set. $\pm$ denotes the standard deviation. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r).

OOD - HateXplain - BERT

DS		100			200			500			14102		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	Ethos	WSF	Dh	Ethos	WSF	Dh	Ethos	WSF	Dh	Ethos	WSF	Dh
None		.21 ± .06	.18 ± .07	.37 ± .03	.16 ± .03	.14 ± .03	.35 ± .04	.32 ± .12	.35 ± .18	.43 ± .03	.57 ± .02	.66 ± .03	.57 ± .02
$\mathcal{L}_{EAR}$		.21 ± .06	.18 ± .07	.37 ± .03	.17 ± .04	.16 ± .05	.35 ± .03	.32 ± .14	.36 ± .2	.44 ± .03	.55 ± .0	.65 ± .01	.56 ± .02
$\mathcal{L}_{AM}$	h	.15 ± .0	.11 ± .0	.35 ± .04	.15 ± .0	.11 ± .0	.35 ± .04	.15 ± .0	.10 ± .0	.35 ± .04	.62 ± .0	.68 ± .0	.60 ± .01
	r	.15 ± .0	.11 ± .0	.35 ± .04	.15 ± .0	.11 ± .0	.35 ± .04	.15 ± .0	.10 ± .0	.35 ± .04	.59 ± .0	.67 ± .0	.58 ± .02
$\mathcal{L}_{KLD}$	h	.15 ± .0	.10 ± .0	.35 ± .04	.14 ± .0	.10 ± .0	.35 ± .04	.17 ± .0	.18 ± .0	.36 ± .05	.61 ± .0	.69 ± .0	.60 ± .02
	r	.15 ± .0	.10 ± .0	.35 ± .04	.19 ± .03	.15 ± .03	.38 ± .06	.24 ± .05	.20 ± .03	.41 ± .04	.58 ± .0	.67 ± .0	.57 ± .03
$\mathcal{L}_{MAE}$	h	.15 ± .0	.10 ± .0	.35 ± .04	.15 ± .0	.10 ± .0	.35 ± .04	.15 ± .0	.11 ± .0	.35 ± .04	.55 ± .0	.67 ± .0	.57 ± .03
	r	.15 ± .0	.10 ± .0	.35 ± .04	.15 ± .0	.10 ± .0	.35 ± .04	.22 ± .07	.20 ± .08	.38 ± .05	.57 ± .01	.67 ± .01	.56 ± .02
$\mathcal{L}_{MSE}$	h	.15 ± .0	.10 ± .0	.35 ± .04	.15 ± .0	.10 ± .0	.35 ± .04	.30 ± .0	.30 ± .0	.43 ± .09	.58 ± .0	.67 ± .0	.59 ± .02
	r	.15 ± .0	.10 ± .0	.35 ± .04	.15 ± .0	.10 ± .0	.35 ± .04	.23 ± .04	.20 ± .04	.38 ± .06	.56 ± .02	.64 ± .01	.56 ± .02
$\mathcal{L}_{OL}$	h	.15 ± .0	.10 ± .0	.35 ± .04	.15 ± .0	.10 ± .0	.35 ± .04	.15 ± .0	.10 ± .0	.35 ± .04	.61 ± .0	.67 ± .0	.58 ± .02
	r	.15 ± .0	.10 ± .0	.35 ± .04	.15 ± .0	.10 ± .0	.35 ± .04	.15 ± .0	.10 ± .0	.35 ± .04	.60 ± .0	.68 ± .0	.58 ± .02

Table 19: Out-of-domain performance for models trained on HateXplain with BERT. We report macro average F1-scores for Ethos, White Supremacist Forum (WSF), and Dynahate (Dh). DS denotes the number of instances in the HateXplain training set. ± denotes the standard deviation. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r).

OOD - VAST - BERT

DS		100		200		519	
$\mathcal{L}_r$	$\boldsymbol{a}^h$	PStance	SemEval6	PStance	SemEval6	PStance	SemEval6
None		.39 ± .07	.25 ± .0	.5 ± .0	.28 ± .0	.49 ± .0	.28 ± .0
$\mathcal{L}_{EAR}$		.39 ± .08	.25 ± .0	.46 ± .0	.24 ± .0	.49 ± .0	.28 ± .0
$\mathcal{L}_{AM}$	h	.37 ± .0	.25 ± .0	.49 ± .0	.25 ± .0	.44 ± .0	.25 ± .0
	r	.37 ± .0	.25 ± .0	.51 ± .02	.26 ± .01	.44 ± .0	.25 ± .0
$\mathcal{L}_{KLD}$	h	.35 ± .0	.25 ± .0	.45 ± .0	.25 ± .0	.44 ± .0	.25 ± .0
	r	.35 ± .0	.25 ± .0	.46 ± .0	.25 ± .0	.44 ± .0	.25 ± .0
$\mathcal{L}_{MAE}$	h	.52 ± .0	.29 ± .0	.53 ± .0	.28 ± .0	.55 ± .0	.36 ± .0
	r	.35 ± .0	.25 ± .0	.46 ± .04	.24 ± .0	.53 ± .03	.3 ± .06
$\mathcal{L}_{MSE}$	h	.46 ± .0	.23 ± .0	.52 ± .0	.3 ± .0	.52 ± .0	.31 ± .0
	r	.36 ± .0	.25 ± .0	.52 ± .0	.3 ± .0	.53 ± .0	.31 ± .0
$\mathcal{L}_{OL}$	h	.44 ± .0	.45 ± .0	.54 ± .0	.44 ± .0	.53 ± .0	.39 ± .0
	r	.46 ± .01	.46 ± .01	.53 ± .02	.32 ± .05	.54 ± .02	.48 ± .01

Table 20: Out-of-domain performance for models trained on VAST with BERT. We report macro average F1-scores for PStance and SemEval6. DS denotes the number of instances in the VAST training set. ± denotes the standard deviation. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r).

OOD - IMDb - DeBERTa

$\mathcal{L}_r$	DS	$\boldsymbol{a}^h$	100				200				500				1472			
			Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU
None			.41 ± .11	.85 ± .01	.81 ± .03	.39 ± .07	.35 ± .01	.76 ± .03	.72 ± .04	.34 ± .01	.46 ± .14	.53 ± .33	.51 ± .3	.43 ± .1	.7 ± .06	.92 ± .01	.87 ± .03	.64 ± .11
$\mathcal{L}_{EAR}$			.41 ± .11	.85 ± .01	.81 ± .03	.39 ± .07	.35 ± .01	.76 ± .03	.72 ± .04	.34 ± .01	.46 ± .14	.53 ± .33	.51 ± .3	.43 ± .1	.62 ± .05	.88 ± .03	.78 ± .11	.5 ± .07
$\mathcal{L}_{AM}$	h		.35 ± .0	.84 ± .03	.82 ± .03	.34 ± .0	.5 ± .04	.8 ± .09	.76 ± .07	.4 ± .02	.55 ± .06	.72 ± .3	.68 ± .3	.47 ± .01	.62 ± .04	.93 ± .01	.86 ± .01	.54 ± .03
	r		.35 ± .0	.84 ± .03	.82 ± .03	.34 ± .0	.5 ± .04	.8 ± .09	.76 ± .07	.4 ± .02	.49 ± .07	.71 ± .29	.68 ± .3	.43 ± .07	.6 ± .07	.92 ± .01	.86 ± .01	.52 ± .07
$\mathcal{L}_{KLD}$	h		.42 ± .11	.86 ± .01	.8 ± .02	.39 ± .08	.46 ± .07	.76 ± .12	.74 ± .08	.39 ± .06	.34 ± .0	.79 ± .03	.78 ± .02	.34 ± .0	.49 ± .1	.86 ± .09	.82 ± .1	.4 ± .06
	r		.38 ± .04	.8 ± .02	.77 ± .05	.36 ± .01	.46 ± .07	.76 ± .12	.74 ± .08	.39 ± .06	.34 ± .0	.79 ± .03	.78 ± .02	.34 ± .0	.43 ± .07	.84 ± .07	.8 ± .09	.37 ± .03
$\mathcal{L}_{MAE}$	h		.42 ± .11	.85 ± .02	.79 ± .01	.39 ± .08	.45 ± .01	.8 ± .03	.74 ± .02	.37 ± .0	.44 ± .04	.89 ± .0	.85 ± .01	.37 ± .02	.51 ± .03	.91 ± .02	.86 ± .04	.42 ± .03
	r		.42 ± .11	.85 ± .02	.79 ± .01	.39 ± .08	.45 ± .01	.8 ± .03	.74 ± .02	.37 ± .0	.41 ± .08	.83 ± .11	.79 ± .09	.37 ± .03	.48 ± .04	.91 ± .02	.85 ± .04	.4 ± .03
$\mathcal{L}_{MSE}$	h		.48 ± .03	.85 ± .06	.8 ± .05	.39 ± .03	.33 ± .0	.74 ± .01	.8 ± .01	.33 ± .0	.41 ± .05	.82 ± .03	.78 ± .03	.35 ± .01	.4 ± .09	.82 ± .03	.82 ± .02	.35 ± .02
	r		.48 ± .03	.85 ± .06	.8 ± .05	.39 ± .03	.33 ± .0	.74 ± .01	.8 ± .01	.33 ± .0	.38 ± .06	.79 ± .03	.78 ± .02	.34 ± .01	.37 ± .05	.83 ± .05	.83 ± .02	.35 ± .02
$\mathcal{L}_{OL}$	h		.33 ± .0	.35 ± .0	.33 ± .0	.33 ± .0	.34 ± .01	.77 ± .01	.74 ± .02	.34 ± .0	.45 ± .04	.89 ± .02	.85 ± .03	.38 ± .02	.35 ± .01	.82 ± .06	.79 ± .06	.35 ± .01
	r		.43 ± .07	.83 ± .02	.78 ± .02	.38 ± .05	.34 ± .01	.77 ± .01	.74 ± .02	.34 ± .0	.49 ± .05	.89 ± .02	.84 ± .06	.41 ± .03	.36 ± .02	.84 ± .04	.8 ± .04	.35 ± .01

Table 21: Out-of-domain performance for models trained on IMDb with DeBERTa. We report macro average F1-scores for Dynasent 1 (Ds1), Dynasent 2 (Ds2), Yelp (Y), and SFU review (SFU). DS denotes the number of instances in the IMDb training set. ± denotes the standard deviation. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r).

OOD - SST - DeBERTa

DS			100				200				500				1472			
$\mathcal{L}_r$	$\boldsymbol{a}^h$		Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU	Ds1	Ds2	Y	SFU
None			.33 ± .0	.32 ± .0	.33 ± .0	.33 ± .0	.74 ± .01	.94 ± .0	.83 ± .03	.77 ± .01	.72 ± .02	.92 ± .02	.77 ± .1	.76 ± .02	.75 ± .01	.92 ± .02	.83 ± .06	.76 ± .01
$\mathcal{L}_{EAR}$			.46 ± .21	.5 ± .32	.48 ± .26	.46 ± .22	.74 ± .01	.94 ± .0	.83 ± .03	.77 ± .01	.72 ± .02	.92 ± .02	.77 ± .1	.76 ± .02	.75 ± .01	.92 ± .02	.83 ± .06	.76 ± .01
$\mathcal{J}_{AM}$	h		.72 ± .01	.9 ± .02	.71 ± .04	.77 ± .01	.74 ± .01	.91 ± .04	.77 ± .05	.76 ± .03	.73 ± .02	.91 ± .04	.78 ± .1	.76 ± .01	.75 ± .0	.94 ± .01	.83 ± .07	.77 ± .01
	r		.72 ± .01	.9 ± .02	.71 ± .04	.77 ± .01	.74 ± .01	.91 ± .04	.77 ± .05	.76 ± .03	.73 ± .02	.91 ± .04	.78 ± .1	.76 ± .01	.75 ± .0	.94 ± .01	.83 ± .07	.77 ± .01
$\mathcal{J}_{KLD}$	h		.33 ± .0	.32 ± .0	.33 ± .0	.33 ± .0	.75 ± .01	.92 ± .01	.77 ± .06	.78 ± .02	.74 ± .01	.94 ± .01	.81 ± .03	.77 ± .01	.75 ± .01	.94 ± .01	.8 ± .05	.78 ± .01
	r		.33 ± .0	.32 ± .0	.33 ± .0	.33 ± .0	.75 ± .01	.92 ± .01	.77 ± .06	.78 ± .02	.74 ± .01	.94 ± .01	.81 ± .03	.77 ± .01	.75 ± .01	.94 ± .01	.8 ± .05	.78 ± .01
$\mathcal{J}_{MAE}$	h		.7 ± .04	.89 ± .03	.77 ± .05	.74 ± .06	.74 ± .01	.94 ± .0	.85 ± .02	.78 ± .01	.73 ± .02	.93 ± .0	.83 ± .02	.75 ± .02	.75 ± .01	.94 ± .0	.84 ± .07	.77 ± .01
	r		.7 ± .04	.89 ± .03	.77 ± .05	.74 ± .06	.74 ± .01	.94 ± .0	.85 ± .02	.78 ± .01	.73 ± .02	.93 ± .0	.83 ± .02	.75 ± .02	.75 ± .01	.94 ± .0	.84 ± .07	.77 ± .01
$\mathcal{J}_{MSE}$	h		.71 ± .04	.91 ± .01	.74 ± .11	.75 ± .03	.75 ± .01	.92 ± .01	.76 ± .06	.77 ± .02	.72 ± .03	.91 ± .05	.79 ± .09	.74 ± .05	.74 ± .02	.95 ± .0	.85 ± .03	.77 ± .01
	r		.46 ± .22	.51 ± .33	.49 ± .27	.48 ± .25	.75 ± .01	.92 ± .01	.76 ± .06	.77 ± .02	.72 ± .03	.91 ± .05	.79 ± .09	.74 ± .05	.74 ± .02	.95 ± .0	.85 ± .03	.77 ± .01
$\mathcal{J}_{OL}$	h		.72 ± .01	.9 ± .01	.64 ± .17	.77 ± .0	.74 ± .02	.92 ± .02	.78 ± .04	.77 ± .02	.74 ± .01	.92 ± .03	.78 ± .05	.77 ± .01	.74 ± .01	.94 ± .02	.77 ± .17	.77 ± .02
	r		.72 ± .01	.9 ± .01	.64 ± .17	.77 ± .0	.74 ± .02	.92 ± .02	.78 ± .04	.77 ± .02	.74 ± .01	.92 ± .03	.78 ± .05	.77 ± .01	.74 ± .01	.94 ± .02	.77 ± .17	.77 ± .02

Table 22: Out-of-domain performance for models trained on SST with DeBERTa. We report macro average F1-scores for Dynasent 1 (Ds1), Dynasent 2 (Ds2), Yelp (Y), and SFU review (SFU). DS denotes the number of instances in the SST training set. ± denotes the standard deviation. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r).

OOD - HateXplain - DeBERTa

DS		100			200			500			14102		
$\mathcal{L}_r$	$\boldsymbol{a}^h$	Ethos	WSF	Dh	Ethos	WSF	Dh	Ethos	WSF	Dh	Ethos	WSF	Dh
None		.14 ± .0	.10 ± .0	.34 ± .04	.14 ± .0	.10 ± .0	.34 ± .04	.39 ± .08	.47 ± .13	.48 ± .04	.33 ± .16	.36 ± .22	.43 ± .01
$\mathcal{L}_{EAR}$		.14 ± .0	.10 ± .0	.34 ± .04	.14 ± .0	.10 ± .0	.34 ± .04	.39 ± .08	.47 ± .13	.48 ± .04	.33 ± .16	.36 ± .22	.43 ± .01
$\mathcal{L}_{AM}$	h	.14 ± .0	.10 ± .0	.34 ± .04	.16 ± .04	.12 ± .03	.35 ± .04	.34 ± .1	.36 ± .16	.44 ± .05	.37 ± .22	.39 ± .3	.46 ± .02
	r	.14 ± .0	.10 ± .0	.34 ± .04	.17 ± .03	.13 ± .03	.35 ± .04	.27 ± .07	.26 ± .09	.40 ± .05	.22 ± .12	.19 ± .16	.38 ± .03
$\mathcal{L}_{KLD}$	h	.14 ± .0	.10 ± .0	.34 ± .04	.14 ± .0	.10 ± .0	.34 ± .04	.30 ± .13	.33 ± .13	.42 ± .03	.38 ± .19	.48 ± .22	.48 ± .02
	r	.14 ± .0	.10 ± .0	.34 ± .04	.15 ± .01	.11 ± .01	.35 ± .04	.34 ± .08	.35 ± .1	.43 ± .03	.38 ± .19	.49 ± .23	.48 ± .01
$\mathcal{L}_{MAE}$	h	.14 ± .0	.10 ± .0	.34 ± .04	.15 ± .02	.11 ± .02	.34 ± .04	.31 ± .06	.30 ± .07	.42 ± .04	.39 ± .21	.50 ± .23	.50 ± .02
	r	.14 ± .0	.10 ± .0	.34 ± .04	.16 ± .02	.12 ± .01	.35 ± .04	.31 ± .06	.30 ± .08	.42 ± .05	.37 ± .17	.48 ± .2	.48 ± .02
$\mathcal{L}_{MSE}$	h	.14 ± .0	.10 ± .0	.34 ± .04	.14 ± .0	.10 ± .0	.34 ± .04	.27 ± .1	.29 ± .11	.40 ± .05	.36 ± .2	.45 ± .3	.48 ± .01
	r	.14 ± .0	.10 ± .0	.34 ± .04	.15 ± .02	.11 ± .02	.34 ± .04	.28 ± .09	.28 ± .12	.41 ± .04	.36 ± .19	.40 ± .24	.42 ± .01
$\mathcal{L}_{OL}$	h	.14 ± .0	.10 ± .0	.34 ± .04	.14 ± .0	.10 ± .0	.34 ± .04	.33 ± .17	.36 ± .25	.44 ± .02	.39 ± .18	.44 ± .29	.48 ± .03
	r	.14 ± .0	.10 ± .0	.34 ± .04	.15 ± .02	.11 ± .01	.34 ± .04	.38 ± .1	.41 ± .17	.46 ± .03	.40 ± .2	.45 ± .29	.44 ± .01

Table 23: Out-of-domain performance for models trained on HateXplain with DeBERTa. We report macro average F1-scores for Ethos, White Supremacist Forum (WSF), and Dynahate (Dh). DS denotes the number of instances in the HateXplain training set. ± denotes the standard deviation. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r).

OOD - VAST - DeBERTa

DS		100		200		519	
$\mathcal{L}_r$	$\boldsymbol{a}^h$	PStance	SemEval6	PStance	SemEval6	PStance	SemEval6
None		.45 ± .17	.52 ± .16	.53 ± .02	.63 ± .0	.36 ± .12	.51 ± .13
$\mathcal{L}_{EAR}$		.45 ± .17	.52 ± .16	.53 ± .02	.63 ± .0	.36 ± .12	.51 ± .13
$\mathcal{L}_{AM}$	h	.48 ± .1	.58 ± .05	.54 ± .02	.6 ± .02	.47 ± .02	.65 ± .02
	r	.48 ± .1	.58 ± .05	.54 ± .02	.6 ± .02	.46 ± .02	.63 ± .02
$\mathcal{L}_{KLD}$	h	.52 ± .05	.6 ± .03	.53 ± .02	.62 ± .01	.43 ± .03	.63 ± .01
	r	.52 ± .05	.6 ± .03	.53 ± .02	.62 ± .01	.43 ± .03	.63 ± .01
$\mathcal{L}_{MAE}$	h	.56 ± .01	.61 ± .01	.46 ± .08	.57 ± .05	.46 ± .01	.64 ± .01
	r	.56 ± .01	.61 ± .01	.46 ± .08	.57 ± .05	.46 ± .01	.64 ± .01
$\mathcal{L}_{MSE}$	h	.47 ± .12	.56 ± .06	.52 ± .01	.63 ± .01	.46 ± .01	.64 ± .01
	r	.47 ± .12	.56 ± .06	.52 ± .01	.63 ± .01	.46 ± .01	.64 ± .01
$\mathcal{L}_{OL}$	h	.55 ± .01	.6 ± .03	.47 ± .09	.58 ± .06	.43 ± .02	.61 ± .02
	r	.55 ± .01	.6 ± .03	.47 ± .09	.58 ± .06	.43 ± .02	.61 ± .02

Table 24: Out-of-domain performance for models trained on VAST with DeBERTa. We report macro average F1-scores for PStance and SemEval6. DS denotes the number of instances in the VAST training set. ± denotes the standard deviation. $\boldsymbol{a}^h$ refers to human-annotated rationale tokens (h) or randomly selected tokens (r).

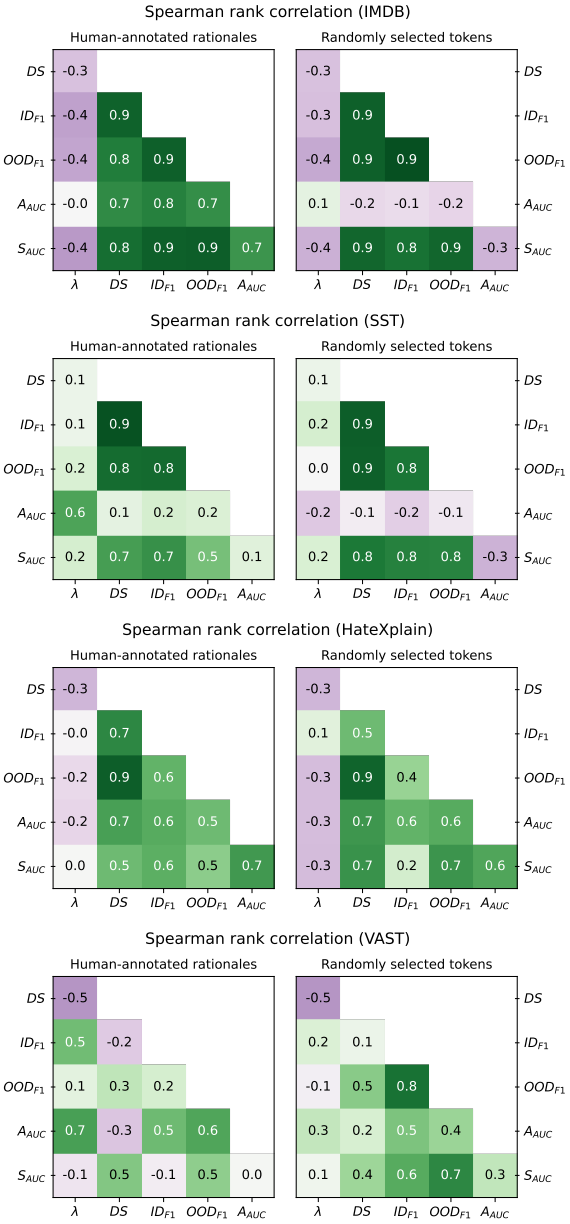


Figure 4: Spearman rank correlation between metrics for BERT models regularised with human-annotated rationale tokens (left) and randomly selected tokens (right). We combine results for $\mathcal{L}_{AM}$, $\mathcal{L}_{KLD}$, $\mathcal{L}_{MAE}$, $\mathcal{L}_{MSE}$, and $\mathcal{L}_{OL}$.