

Learning from *Sufficient* Rationales: Analysing the Relationship Between Explanation Faithfulness and Token-level Regularisation Strategies

Jonathan Kamp¹

Lisa Beinborn²

Antske Fokkens¹

¹Computational Linguistics and Text Mining Lab, Vrije Universiteit Amsterdam

²Institute of Computer Science, University of Göttingen

{j.b.kamp, antske.fokkens}@vu.nl, lisa.beinborn@uni-goettingen.de

Abstract

Human explanations of natural language, *rationales*, form a tool to assess whether models learn a label *for the right reasons* or rely on dataset-specific shortcuts. *Sufficiency* is a common metric for estimating the informativeness of rationales, but it provides limited insight into the effects of rationale information on model performance. We address this limitation by relating sufficiency to two modelling paradigms: the ability of models to identify which tokens are part of the rationale (through token classification) and the ability of improving model performance by incorporating rationales in the input (through attention regularisation). We find that highly informative rationales are not likely to help classify the instance correctly. Sufficiency conversely captures the classification impact of the non-rationalised context, which interferes with rationale information in the same input. We also find that incorporating rationale information in model inputs can boost cross-domain classification, but results are inconsistent per task and model type. Finally, sufficiency and token classification appear to be unrelated. These results exemplify the complexity of rationales, showing that metrics capable of systematically capturing this type of information merit further investigation.

1 Introduction

Neural text classifiers are trained to generalise patterns in texts to predict labels for unseen examples. The learnt patterns optimise the classification objective but do not necessarily align with a human interpretation of the task (Jakobsen et al., 2023; Ross et al., 2017). For example, an argument classification model tends to overfit to topic-specific vocabulary instead of capturing general argumentation structure. To uncover such over-simplified classification patterns, we can compare model behaviour to human-annotated reasons for input labels, i.e. *rationales* (Carton et al., 2020; Strout et al., 2019).

	Input tokens	$\mathcal{M}(\cdot)_+$
x	One of the greatest family - oriented , fantasy - adventure movies ever .	0.6
r	the greatest family - oriented	0.8

Figure 1: We compare the model \mathcal{M} probability of the gold label $+$ when using the full input x to the probability on the isolated rationales r . The example is an instance from the sentiment analysis dataset SST (Socher et al., 2013; Carton et al., 2020).

When provided in *highlighted* form (Wiegrefe and Marasović, 2021) rather than e.g. free-text, token-level rationales establish a benchmark for analysing model behaviour at high granularity patterns within the input, such as words and phrases.

Sufficiency (Lei et al., 2016) is commonly adopted to quantify the contribution of highlight rationales to model predictions, i.e., *rationale informativeness*, but the interpretation of this metric remains opaque. In principle, sufficiency corresponds to the change in model confidence when reducing the input to the rationale. Its use is prevalent as a *faithfulness* proxy for post-hoc attribution methods (Manna and Sett, 2024; Kamahi and Yaghoobzadeh, 2024), automated techniques that approximate each input token’s informativeness to the model’s prediction. By assessing their informativeness, sufficiency measures how accurately these artificial token-level rationales truly reflect the model’s inner workings. In Figure 1, the probability for positive sentiment is higher for rationales r ("the greatest family-oriented") compared to input x , suggesting that, in this case, the non-rationale words constitute noise that steers the probability away from positive sentiment. However, the current use of sufficiency is limited to a *relative* faithfulness estimation when ranking attribution methods, while *high* and *low* sufficiency scores remain vague in the absolute sense. In fact, such use (i) oversimplifies the interaction between important and (seemingly less relevant) contextual linguistic

cues in the input, and (ii) does not provide guidance into how rationale informativeness can be exploited by a model; we address this limitation.

We investigate the role that sufficiency plays in capturing rationale informativeness by establishing its relation to two modelling paradigms with rationale inputs. In particular, we investigate the relation between sufficiency and (a) the ability of a model to identify tokens that are part of the rationale and (b) the ability of rationales to improve a model’s performance. The way in which these paradigms profit from rationales is expected to reflect how sufficiency captures rationale informativeness. Although input rationalisation has been shown to improve both in- and cross-domain classification (Hartmann and Sonntag, 2022), under which conditions rationales are helpful remains an open question (Hase and Bansal, 2022). In addition to exploring the relation to sufficiency, this paper is, to our knowledge, also first to systematically assess the effect of adding rationale information to input across tasks, models and learning strategies. We have mixed results, though our BERT model benefits from rationales in a cross-domain setting.

Our results do not confirm the hypothesis that rationales that are highly informative (based on their sufficiency score) consist of easily identifiable tokens, nor the hypothesis that good sufficiency scores predict that rationales improve classification. Instead, our research reveals that sufficiency captures the impact of non-rationale contexts on model predictions. This highlights that the relation between rationales and their context is complex and that sufficiency by itself can only capture relative information.¹

2 Related Work

We first describe sufficiency’s role as a key metric for model understanding (§2.1) and then describe how rationales have been used to improve model performance in previous work (§2.2).

2.1 Sufficiency for Faithfulness Assessment

Token-level explanations are subject to inter-rater variability, both when the rater is human or artificial. When explanations are extracted artificially from a fine-tuned model, e.g. via saliency methods, assessing their *faithfulness* to the inner decision making processes is crucial to model interpretability.

¹We publicly release our code at <https://github.com/jbkamp/repo-Suff-Rationales>.

ity. This is especially relevant given the tendency of explanations to disagree, even on the same instance (Neely et al., 2022; Kamp et al., 2024). Research on faithfulness assessment keeps being relevant; for example, Fayyaz et al. (2024) show the importance of faithfulness evaluation for Large Language Models (LLMs), finding that feature attributions are more faithful than prompt-based self-explanations and align better with rationales. Madsen et al. (2024a) propose consistency checks for self-explanations. A question regarding human explanations, instead, is whether they are informative enough for a model to learn from. *Sufficiency* (Lei et al., 2016) serves as a key metric for both human and model explanations.

In line with other faithfulness metrics, sufficiency is ablation-based. In DeYoung et al. (2020), collected rationales are assessed on sufficiency based on the notion of contrast examples (Zaidan et al., 2007): rationales are isolated from their context, resulting in a change of probability towards the target class of a test instance. Conversely, *comprehensiveness* (Yu et al., 2019) measures the change in probability by removing the rationales. Similarly, Hooker et al. (2019) observe the effect of removing important features by training different models. We investigate sufficiency (formalised in §4) as it may indirectly estimate the importance of both rationale words and context words.

Calculating sufficiency may lead to out-of-distribution examples when non-rationales are removed or masked (Madsen et al., 2024b; Carton et al., 2022; Hooker et al., 2019). In fact, when we do so, we often end up with incomplete and ungrammatical sentences causing irregular model behaviour. A handful of studies investigate the difference between removing and masking non-rationales on a single model type (Kamahi and Yaghoobzadeh, 2024; Carton et al., 2022). We find no differences for models with different pre-training objectives.

The unique contribution of our work is that we explicitly link sufficiency to model improvement, investigating both masked language models and autoregressive decoders. We bridge the gap from rationale characteristics, which we specify as informativeness to a prediction, to learnability.

2.2 Learning from Rationales

The power of deep learning lies in learning patterns in the data without explicit instructions. A commonly undesired consequence is the tendency

for models to learn dataset-specific shortcuts. Human rationales can be used as an additional source of task-specific information to steer the learning process. Hase and Bansal (2022) compare different types of rationales and find that rationales as highlights are more beneficial for improving fine-tuned performance of a retrieval task than free-text explanations. They experiment with a synthetic task showing that retrieval-based RoBERTa models (Liu et al., 2019; Reimers and Gurevych, 2019) improve with explanations encoded in the input. Recent work on learning from rationales has been implemented on BERT-like models through attention regularisation. The classifier jointly computes the main loss term and a second term optimising the attention weights to a given input rationale mask, often provided by feature attribution methods. Results from previous studies are heterogeneous. Some show that additional rationale information is beneficial to the task: for example, Attanasio et al. (2022) find that rationalisation mitigates unintended gender bias in hate speech detection and favours cross-domain generalisation; Stacey et al. (2022) find improvements on both in- and cross-domain natural language inference; Carton et al. (2022) find accuracy improvements between 0.4% and 3% on three different in-domain tasks; Krishna et al. (2023) and Bhan et al. (2024) find that attribution methods’ explanations (fed back as additional prompt context) can improve instruction-tuned LLMs such as Mistral-7B (Jiang et al., 2023). In contrast, Ferreira et al. (2025) find that guiding the model with rationales does not necessarily improve on an out-of-domain classification task; Kozlova et al. (2024) exploit eyetracking data to guide encoder-only and encoder-decoder models on an anaphora resolution task, finding mostly no improvement on the baseline models.

In a similar fashion, Su et al. (2024) adopt attention regularisation in pre-training to guide a transformer when crucial cues (comments, function names) to code understanding are absent. In argument mining, Jayaram and Allaway (2021) fine-tune a stance predictor and use rationales as priors to optimise MAW, a feature attribution method based on the model’s attention weights, in producing more human-like explanations. While they show that MAW is relatively *faithful* compared to Gradient \times Input (Shrikumar et al., 2016) and a random baseline, the link between rationale faithfulness and informativeness remains mostly unclear. Comparably, Xie et al. (2024) fine-tune Electra

(Clark et al., 2020), Llama-2-7b (Touvron et al., 2023), and GPT-2-medium (Radford et al., 2019) to produce explanations by regularising the attention distribution via different loss terms that optimise for faithfulness criteria (such as *sufficiency*). They outperform other types of model guidance. This shows that sufficiency can be a beneficial component for learning through attention regularisation, which we will explore further.

Hartmann and Sonntag (2022) train a token classifier simultaneously with a regular classifier (in a multi-task learning fashion) to produce explanations for the main task. Ross et al. (2022) find that this technique, also called *self-rationalisation*, produces task- and model-specific effects. Pruthi et al. (2022) find that attention regularisation outperforms the multi-task learning approach; this suggests that either the token classifier is inadequate to learn enough from the input rationales, or the learnt information is not fully utilised through the shared parameters of the main task. Carton et al. (2020) evaluate rationale quality; however, the link between rationale characteristics and the ability of a token classifier to distinguish rationales from non-rationales in a sequence is currently underdefined. To our knowledge, different approaches that use rationales for model improvement have not yet been systematically compared for a wider range of tasks.

3 Data and Models

We describe the classification tasks and model types for which we compute sufficiency (expected to quantify rationale information) and for which we assess the ability to learn from rationales. By combining these two aspects in later sections, we aim to understand the way in which sufficiency captures rationale informativeness.

3.1 Data

The inputs for our models are *rationalised inputs*: tokenised texts with an accompanying binary mask indicating whether a token is part (1) or not part (0) of the rationale. Sets of rationalised inputs can be created in two ways. (i) Rationales are added to labeled sentences (or larger units), e.g. e-SNLI (Camburu et al., 2018) and HateXplain (Mathew et al., 2021). (ii) Token-level labels are repurposed (Wiegrefe and Marasović, 2021) such as in AURC-8 (Trautmann et al., 2020), where sequences of tokens labeled as argument units from the original dataset become the rationales for the

new task of predicting whether a sentence contains an argument. We aim for a representative sample of tasks selecting four existing datasets and adding two new repurposed ones. They are presented in Table 1 and described below.

Dataset	#cl	gran	type	$ r / x $
AURC-8	2	None	Semantic	.41
SST	2	None	Semantic	.46
HateXplain	3	Phrase	Semantic	.19
e-SNLI	3	Word	Both	.21
CoNLL-chunk	2	Phrase	Syntactic	.01
CoNLL-NER	2	None	Syntactic	.07

Table 1: The datasets vary in: **#classes** (2 or 3); the **granularity** restriction in the annotations (none, word- and phrase-level, following Wiegrefe and Marasović (2021)); the **type** of linguistic cues the model is expected to mostly rely on (syntactic or semantic, following Jang et al. (2024)); the rationale density: the average number of rationale tokens over input tokens ($|r|/|x|$).

AURC-8 The texts from Trautmann et al. (2020) are labelled as *argumentative* or *non-argumentative* in the context of a given topic or *domain* (e.g. minimum wage). We explore both original topic-based splits of the data: in-domain (AURC-8_{ID}; 4,193 training and 1,200 test instances on same topics) and cross-domain (AURC-8_{CD}; 4,000 training and 2,000 test instances on different topics).

Stanford Sentiment Treebank (SST) Following Carton et al. (2020), we obtain the rationales from SST (Socher et al., 2013) by flattening the original syntactic phrase-level annotations of movie review snippets (see Appendix A.2 for details). The labels are positive and negative. There are 6,917 training and 1,821 test instances.

HateXplain Texts in this set are labelled as *hate-speech*, *offensive*, *normal* (Mathew et al., 2021). For instances with multiple annotators, we aggregate the labels by majority vote (in line with the authors) and the rationales by taking the union (\cup) or intersection (\cap). We aggregate rationale annotations to obtain binary rationale masks over input tokens, in line with the other datasets that we cover. HateXplain \cup allows for human variation, while HateXplain \cap restricts the rationale to the tokens that were highlighted by all annotators. Each version has 15,379 training and 1,924 test instances.

e-SNLI This dataset by Camburu et al. (2018) contains highlight rationales to explain a natural language inference task. The labels (entailment,

contradiction, neutral) indicate the semantic relationship between two sentences that are concatenated by a [SEP] token. It contains 549,339 training and 9,824 test instances.

CoNLL-chunk We re-purpose the CoNLL-2000 chunking dataset (Tjong Kim Sang and Buchholz, 2000). For each type of syntactic phrase, we create a version of the dataset where the task is to predict its presence or absence. We ignore the three most frequent tags (NP, VP, PP) as they would lead to barely any 0-class sentence labels. We choose the next-most frequent tags with at least 500 occurrences: ADVP, ADJP, SBAR, PRT. Each version (same texts, different labels) has 8,936 training and 2,012 test instances.

CoNLL-NER We re-purpose the CoNLL-2003 NER dataset (Tjong Kim Sang and De Meulder, 2003). For each of the tags *person* (PER), *organisation* (ORG), *location* (LOC), *miscellaneous* (MISC), we create a version of the dataset where the task is to predict its presence or absence. Each has 14,041 training and 3,453 test instances.

3.2 Models

For each classification task, we fine-tune four transformer models that have a comparable number of parameters: **BERT**-base-uncased, 110M (Devlin et al., 2019), **Pythia**-160M (Biderman et al., 2023), **ModernBERT**-base, 149M (Warner et al., 2024), **GPT-Neo**-125M (Black et al., 2021). Each is fine-tuned three times per task with three different random seeds. Similarly, Wyatte et al. (2024) compare bert-large, 330M to Pythia, 410M, which also fall into the same sizes. Among the 70M and 160M Pythia we opt for the latter, being the more popular² middle-sized model of the two. The runtimes of the four models fall in comparable ranges (Appendix A.2). BERT and ModernBERT use bidirectional attention and are therefore expected to perform better than our left-to-right attention models on the classification tasks, although mixed findings emerge from previous work (Bouchiha et al., 2025; Pilicita and Barra, 2025; Lukito et al., 2024).

4 Sufficiency as Contextual Impact

To better understand the added value of rationale information and the role of sufficiency as rationale informativeness proxy, we first formalise sufficiency and compute a corresponding value for each

²146k vs. 120k downloads until December 2024.

instance in the different datasets. We calculate sufficiency as the confidence change of the model \mathcal{M} , by measuring the change in probability between the prediction on instance x_i and its counterexample r_i for a given class j , in line with DeYoung et al. (2020) (see example in Figure 1):

$$\text{suff}(x_i) = \mathcal{M}(x_i)_j - \mathcal{M}(r_i)_j \quad (1)$$

We average sufficiency over all instances in a dataset (D) for a specific model \mathcal{M} :

$$\text{suff}(D) = \frac{1}{n} \sum_{i=1}^n (\mathcal{M}(x_i)_j - \mathcal{M}(r_i)_j) \quad (2)$$

Here, r_i corresponds to x_i in which the non-rationale context tokens are ablated. For context ablation, we apply removal or masking; we mask by using the pre-trained [MASK] token for BERT and ModernBERT and a random embedding for Pythia and GPT-Neo (details in Appendix A.2). As the two implementations produce similar scores (Appendix A.1), we will only cover removal in our analyses. Sufficiency has been defined such that a value of zero indicates that the rationale alone is strictly sufficient to predict the correct label. Higher values indicate that the other context tokens also contribute important information to the prediction. We find it more intuitive to interpret the metric as an indicator for *contextual impact* (CI), i.e., if the value approaches zero, the context does not have additional impact on the prediction (compared to the rationale). High CI is expected to entail low rationale informativeness.

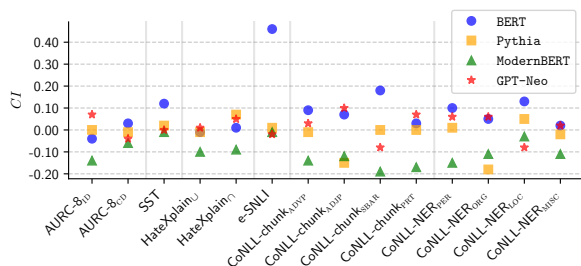


Figure 2: Dataset-average contextual impact (CI).

Figure 2 provides an overview of the dataset-average CI for different tasks and models. For example, the large difference between CI for BERT and the other models on e-SNLI may show that different models encode the balance between context and rationale information differently even on the same data. The overall positive CI by BERT

across tasks might indicate that the model relies more on context information than e.g. ModernBERT, a model characterised by an overall negative CI . This difference might stem from a different model behaviour on missing context (e.g. linguistic connectors) when predicting $\mathcal{M}(r_i)_j$ in Equation 1, where r_i is essentially a bag-of-rationales with preserved order from the source sentence. In §5, we investigate to what extent CI captures rationale informativeness for learning these tasks.

5 Learning from Rationale Information

We investigate the model’s ability to learn from rationales on the six datasets using two learning paradigms: token classification (with metric TC) and attention regularisation (with metric AR). The first paradigm investigates the ability of models to distinguish rationale tokens from the adjacent context. The second paradigm investigates whether including information from rationales in the input improves the model’s performance.

5.1 Learnability Metrics

We quantify the added value of rationale information for our two learning paradigms. Both TC and AR are ratios indicating the performance over a baseline. Compared to absolute differences, ratios make relative improvements interpretable and comparisons meaningful across tasks.

Learnability metric TC We train a binary token classifier \mathcal{T} on rationales. TC is defined as:

$$TC = \text{token-f1}_{\mathcal{T}(D)} / \text{token-f1}_{\mathcal{B}(D)} \quad (3)$$

It measures the relative performance of the token classifier \mathcal{T} on rationales as token labels (computed at the instance-level and averaged over dataset D) compared to a dataset-specific baseline function \mathcal{B} that assigns the majority label to each token. We use this simple baseline as a proxy to contextualise model performance within task-specific complexity (not for benchmarking purposes) making cross-task comparisons more meaningful. A relative score also accounts for different baseline \mathcal{B} performance across models since token labels differ due to different tokenisations. The greater TC , for $TC > 1$, the greater the improvement. We expect a high TC for informative rationales, i.e., for rationales where CI is low.

Learnability metric AR We fine-tune a sequence classifier on inputs that are rationalised during training only, through attention regularisation. AR is

computed as:

$$AR = f1_{\mathcal{R}(D)} / f1_{\mathcal{M}(D)} \quad (4)$$

It represents the relative performance of the regularised model \mathcal{R} with respect to the baseline model \mathcal{M} on the non-rationalised test set. The higher AR , for $AR > 1$, the more the rationale regularization improved the model \mathcal{M} .

The input of \mathcal{R} during training is a tokenised sentence with a sentence-wise label y and a binary rationale mask a (1 for important tokens, else 0) serving as ground truth for guiding the attention mechanism. During the forward pass, we extract the attention weights \hat{a} . The regularised model is then optimised with a second loss term $\mathcal{L}_{\text{attention}}$ (binary cross-entropy) measuring the distance between rationale mask a and attention weights \hat{a} :

$$\mathcal{L} = \mathcal{L}_{\text{task}}(y, \hat{y}) + \mathcal{L}_{\text{attention}}(a, \hat{a}) \quad (5)$$

Kozlova et al. (2024) explored the attention weights from the first, last and a specific layer for which human eye-tracking attention and model correlated most. Due to the extensive fine-tuning, we confine our experiments to the attention values of the last layer as it is closest to the final predictions.

When we isolate the effect of attention regularisation with rationales, we aim to understand the rationales’ informativeness in guiding the model and assess whether regularisation is a useful technique. If we do not observe performance improvements, this might be due to the rationale being redundant (the model would already rely on similar cues) or detrimental (steering the model away from dataset-specific shortcuts that boosted original results). Attention regularisation might optimise better for BERT and ModernBERT (than for Pythia and GPT-Neo) because of the ability to attend to all tokens rather than only the past tokens in the input.

5.2 Results of the Two Learning Paradigms

We describe the effects of attention regularisation through AR and the ability of classifying rationale tokens through TC . We summarise three main observations based on Figure 3. More details are provided in Appendix A.3, including the absolute f1 scores of the models and baselines (Figure 9, Table 6) and results for significance tests (Table 7).

Regularisation is task- and model-dependent

The AR results in Figure 3 show that attention regularisation has a three-run-average positive effect in 6/14 tasks for BERT, 7 for Pythia, 8 for

ModernBERT, and 6 for GPT-Neo. To determine the consistency of observed improvements, we compute 95% bootstrap confidence intervals by extracting the averages of three samples with replacement (1,000 iterations) from the performances of the regularised and baseline models, for each model \times task setup. Given the small sample size of 3 runs per setup, these confidence intervals must be interpreted with caution. We find that the lower bounds of the intervals indicate consistent improvement in 83% of cases for BERT, even for relatively low AR , e.g. < 1.05 . Consistent improvements are lower for the other three models: 43%, 25% and 0%. The latter (GPT-Neo) can be explained by relatively low AR overall, peaking at 1.04. These partly divergent results show that the effectivity of attention regularisation is, besides the expected differences among tasks, highly model dependent. We also find that regularisation does not guarantee a greater stability for predictions on the same instances between models, contrary to the expectation that rationalisation would mitigate the differences between random seeds (Appendix A.1).

Rationales can boost X-domain performance

We take a closer look at in- vs. cross-domain argument mining (AURC-8_{ID}, AURC-8_{CD}). We partly validate the idea that attention regularisation improves performance in both in- and cross-domain settings (Hartmann and Sonntag, 2022). In this specific task, only BERT benefits from attention regularisation on both settings, with cross-domain performance gaining a substantially greater improvement than in-domain ($AR = 1.14$ vs. $AR = 1.01$); the former problem being inherently more difficult appears to be compensated by explicit guidance. Notably, by enhancing sequence classifiers (comparable to the ones by Trautmann et al. (2020)) with a simple loss term and the available rationale data that they use for token classification, we drastically reduce the gap between cross-domain and in-domain argument mining. At the same time, ModernBERT only improves on average but non-consistently in the in-domain setting ($AR = 1.01$, [0.99, 1.02]), while the two autoregressive models do not benefit from rationale-based regularisation.

ModernBERT > autoregressive models on TC

We expected that our masked language models would perform best on token classification. While TC is higher for Pythia and GPT-Neo on 9 out of 14 tasks compared to BERT (Figure 3), ModernBERT outperforms the former two on at least 10 tasks.

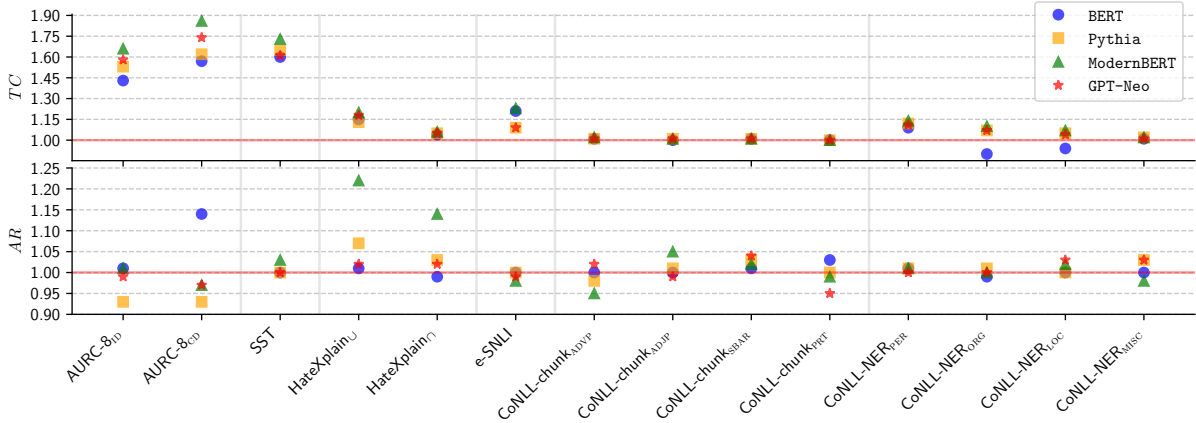


Figure 3: Performance results expressed in TC and AR . Scores > 1.00 indicate model improvement over the baselines described in §5.1. For example, 1.15 indicates a relative improvement by a factor of 1.15, or +15%.

Still, no strong model generalisations can be drawn from this small sample size. Interestingly, the classifiers with the highest TC are the ones that were trained on the datasets with highest rationale density $|r|/|x|$. This is likely because the majority class baseline is higher for tasks where the rationales are less dense (Table 6, Appendix A.3).

5.3 Relating Dataset- CI to Learnability

The relation between CI and learning from rationales remains unclear: what is it that CI captures about rationales that a model can use? Hence, the question we address is whether CI aligns with the learnability metrics TC and AR .

As a first exploratory step, we analyse the relation between dataset-average CI (Figure 2) and performance. We therefore compute pairwise correlations on (CI, TC) , (CI, AR) . The learnability metrics $(\frac{f1}{bl})$, where bl is the baseline) are first normalised as $(\frac{f1}{bl} \cdot bl) - bl / (1 - bl)$ to reflect the relative improvement to both the baseline and the upper performance limit (1.0 $f1$), giving a more balanced weight to the improvement of models that have a high base performance. Given the small sample size (6 condensed task versions to ensure statistical independence) the correlation’s magnitude may be unstable but can still provide indications of polarity. We adopt Spearman’s r because of its higher outlier robustness compared to e.g. Pearson’s r . We hypothesise negative correlations between dataset-level CI and performance based on the intuition that low CI entails high rationale informativeness, hence stronger cues for guiding the model.

CI might correlate with AR for BERT Cautiously in line with our expectations, we find negative cor-

relations that are moderate for AR with BERT, yet results are not significant ($r = -.49, p = .33$). We observe non-significant absent to weak correlations for the other models and no correlation between CI and TC . As these signs are inconclusive, we extend this small dataset-level analysis with a full instance-level analysis (§5.4), aiming to solidify the understanding of CI and its ties to model performance.

5.4 Relating Instance- CI to Learnability

If low CI really entails high rationale informativeness, we would expect CI to correlate (again, negatively) with prediction correctness of the inputs. Furthermore, we expect that instances with low CI either tend to receive (i) correct predictions by base model \mathcal{M} or (ii) initial incorrect predictions by \mathcal{M} that are flipped by the regularised model \mathcal{R} . We investigate this by analysing the pairs of tasks with respectively the highest and lowest performance increase (AR).³ We expect the relation between CI and predictions to be stronger in tasks for which \mathcal{R} successfully learnt from the rationales (high AR) than in tasks where attention regularisation had a non-positive effect (low AR). We cannot compare models directly to one another as the selected tasks vary; instead, we inspect them in parallel.

No negative correlation, but sometimes positive Figure 4 shows Pearson’s correlation between instance-level CI (continuous scores) and predictions. Prediction correctness is binary⁴ for

³BERT: AURC-8_{CD} & HateXplain_∩; Pythia: HateXplain_∪ & AURC-8_{ID}; ModernBERT: HateXplain_∪ & CoNLL-chunk_{ADVP}; GPT-NeO: CoNLL-chunk_{SBAR} & CoNLL-chunk_{PRT}.

⁴In this case, we compute the point-biserial correlation, which is equivalent to Pearson’s r for continuous variables.

the sequence classifiers \mathcal{M} and \mathcal{R} ; predictions by token classifier \mathcal{T} are continuous f1-scores based on token label correctness. For BERT, contrary to our expectations, the correlation with base model \mathcal{M} and regularised model \mathcal{R} is weak to moderately *positive* for the high AR task, and weak positive to absent for the low AR task. We observe the same drop for \mathcal{T} , but the correlation remains absent. For Pythia, we see a different result: \mathcal{T} predictions still do not correlate, but neither do \mathcal{M} and \mathcal{R} (weak to absent, mostly non-significant). While no correlation is detected for ModernBERT, up to strong positive correlations are found for GPT-Neo, but with no clear differences between high and low AR tasks.

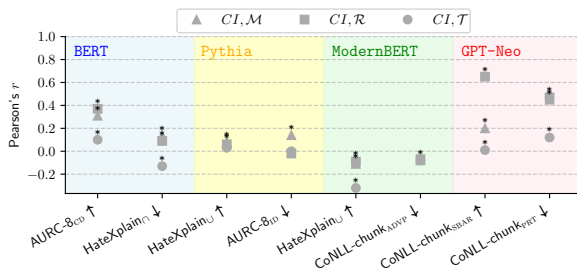


Figure 4: The correlation between CI and predictions in high AR (\uparrow) and low AR (\downarrow) tasks tends to positive, not negative, suggesting that high CI does not entail rationale informativeness. Significance (*) at $p < .05$.

CI meaning becomes clearer at the extremes

Correlations may be mitigated by the most common values (the middle of the CI distribution). In Figure 5, we thus inspect the extremes: instances with bottom and top 10% CI where the link with predictions should be more pronounced. We find that the top 10% CI instances are tied to overall better predictions than the bottom 10%, strengthening the signs of positive correlation (Figure 4).

This tendency for *high* CI implying better predictions stands in contrast to our initial hypotheses. If we update our hypotheses based on this inverse tendency, we would accordingly expect top CI instances to receive better predictions than bottom CI instances when zooming in on the regularised model \mathcal{R} . \mathcal{R} is of interest as it allows us to differentiate between models that did and did not learn from the rationales. Specifically, we expect $\Delta Pred$, the prediction performance distance between top and bottom CI instances, to be stronger in high AR tasks (\uparrow), i.e. tasks where rationalisation had a positive effect, compared to low AR tasks (\downarrow).

Based on Figure 5, we confirm this expectation in all four scenarios: for BERT ($\Delta Pred \uparrow = 62 > \Delta Pred \downarrow = 15$) and GPT-Neo ($75 > 38$), $\Delta Pred$ is clearly stronger in the $\uparrow AR$ task. For Pythia, the expected contrast between $\uparrow AR$ and $\downarrow AR$ tasks is present but weaker ($9 > 4$). For ModernBERT, the distance between top and bottom is marginal in the $\uparrow AR$ task ($\Delta Pred \uparrow = 1$), while the prediction performance for \mathcal{R} on top is even below bottom in the $\downarrow AR$ task ($\Delta Pred \downarrow = -7$). More details about the distances are given in Table 8, Appendix A.3.

By comparing the regularised model \mathcal{R} and base model \mathcal{M} , we finally find that attention regularisation improves predictions especially on high CI instances for models that successfully learnt from rationales ($\uparrow AR$ tasks), except for ModernBERT.

In summary, the fact that *high* instance CI leans to *correct* predictions seems to indicate that CI does not capture rationale informativeness in the way we originally hypothesised. This finding also counters the weak correlation direction found at the dataset level in §5.3. These instance-level results suggest that *low* CI does not tell much about the absolute informativeness of the rationale (to improve a model). Rather, CI appears to capture information about the way in which rationales and non-rationale contexts interact with one another. At least at the extremes of the CI spectrum, *high* CI captures the impact of the context information on the prediction, i.e. in the form of interference with rationale information in the same input.

6 CI and Rationale Aggregation

Since rationale quality is assumed to be central to model improvement, we included two versions of HateXplain (taking advantage of the multiple annotators per instance) to test the effect of different rationale aggregation strategies on model performance and CI . Do rather few, strong rationale words or many, noisy rationale words guide the model better? In theory, for CI to be low, rationales should include key words that are informative about the sentence label and exclude context words that could bias the prediction. Rationale annotation is subject to human disagreement, which arguably defines a rationale's informativeness to some degree: by aggregating rationale annotations through union, these unified rationales would include relatively many words, part of which would have a disputed informativeness ($0 < \text{annotators} < n$ highlighted the word). In contrast, rationales ag-

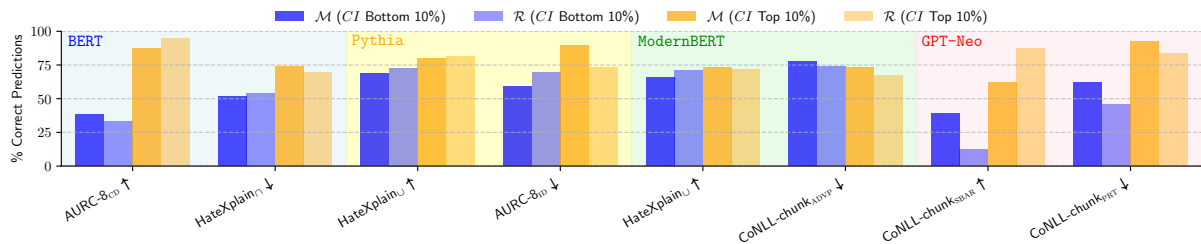


Figure 5: % Correct predictions for bottom and top Contextual Impact. For each model, we compare the respective tasks that achieved highest AR (\uparrow)—where a positive rationalisation effect was observed, vs. lowest AR (\downarrow). Overall, high CI leans to correct predictions. A greater top–bottom distance for \mathcal{R} is observed in high AR (\uparrow) tasks.

gregated by intersection would strictly include key words (where all n annotators agreed on), but lack those disputed words that might carry partial information about the target class. Figure 2 shows that CI is lower for $HateXplain_{\cup}$ than for \neg_{\cap} (for all four models), with also TC and AR being mostly higher (cautiously in line with our negative correlation hypothesis from §5.3).

This suggests that the incorporation of disputed rationales in the input benefits modelling, at least in this scenario of hate speech detection. In other words, collecting different human perspectives on word importance (aggregated through union) is preferred over a single annotator that might label too strictly. In cases where a single annotator per instance is available, the degree to which they act as a *lumper* is thus likely to affect the quality of the collected rationales. In combination with the granularity restrictions given in the annotation instructions, this annotator characteristic potentially influences CI and the learning ability of the model.

7 Conclusion

We carried out a rich set of experiments investigating the role of sufficiency, reframed as CI , in capturing rationale informativeness for model improvement on two different learning paradigms. What is important information and is there a link in the way models process it? Well, it depends: (A) There are signs, although thin, that low dataset-level CI encodes high rationale informativeness for regularised BERT models, cautiously in line with our initial hypotheses. (B) However, our instance-level analyses yield a pattern that on the one hand is clearer but also goes against expectations, pointing to *higher CI, better predictions* (visible at the extremes of the CI spectrum). This suggests that high CI inputs rely on the informativeness stemming from *both* rationales and non-rationale con-

texts. In other words, high CI would entail the context words’ *relative weakness* (in interfering with rationale information in the same input), rather than reflecting in any way the rationales’ *absolute strength*. To this end, *contextual impact* appropriately renames sufficiency. Furthermore, (C) there is no one-size-fits-all solution on how to rationalise a model because information value is strongly determined by the task and by the processor (the model using the rationale or not, the humans disagreeing on its length or content). These heterogeneous results show that, overall, sufficiency is not a good indicator of how to prioritise information in training a model. Within this heterogeneity, though, (D) attention regularisation using rationales shows potential as it closes the gap between in- and cross-domain argument mining performance for BERT by adopting a simple auxiliary loss. (E) Finally, the learnability of a token classifier does not correlate with CI . This surprising finding shows that the relation between rationales and their context is complex and sufficiency can mask underlying processes, e.g. models that perform well on test data due to shortcuts in the context can have high CI despite informative rationales.

Overall, results show that the informativeness of rationales merits further investigation. We would like to extend our analyses to more datasets including other languages (following Kozlova et al. (2024) on Russian). Investigating datasets in different languages has the additional advantage that it can provide insight into both language- and culture-specific components in human perception of sufficient information, which especially arises in unconstrained rationales. Finally, new insights about contextual information can lead to reconsidering *comprehensiveness* (§2.1), i.e. faithfulness through rationale ablation, to further investigate the balance between rationale and context information.

Limitations

As different models are pre-trained using different tokenisers, the alignment between (sub)tokens and rationales may also differ. When comparing models that are pre-trained differently, there is no straightforward solution. Secondly, default hyperparameters were used with only slight adaptations. Keeping hyperparameters mostly constant (i) reduces their confounding impact and (ii) diminishes the ecological footprint and costs. With more extensive hyperparameter tuning, we would achieve more optimally fine-tuned models for the different tasks and datasets. Although we remained cautious in our claims, there always remains a risk that factors such as the above interfere with the interpretation of our model comparisons.

Acknowledgements

This research is funded by the Dutch National Science Organisation (NWO) through project code NWA.1292.19.399 – *InDeep: Interpreting Deep Learning Models for Text and Sound*. Lisa Beinborn’s research is supported by an *Impulsprofessur* grant from the *zukunft.niedersachsen* program and by a VENI grant (VI.Veni.211C.039) from the Dutch National Science Organisation (NWO). We thank the anonymous reviewers for their constructive comments that helped improve the paper. Remaining errors are our own.

References

- 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.
- Milan Bhan, Jean-Noël Vittaut, Nicolas Chesneau, and Marie-Jeanne Lesot. 2024. [Self-AMPLIFY: Improving small language models with self post hoc explanations](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10974–10991, Miami, Florida, USA. Association for Computational Linguistics.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.
- Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. [GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow](#). If you use this software, please cite it using these metadata.
- Djelloul Bouchiha, Abdelghani Bouziane, Nouredine Doumi, Benamar Hamzaoui, and Sofiane Boukli-Hacene. 2025. Hierarchical text classification:: Fine-tuned gpt-2 vs bert-bilstm. *Applied Computer Systems*, 30(1):40–46.
- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.
- Samuel Carton, Surya Kanoria, and Chenhao Tan. 2022. What to learn, and how: Toward effective learning from rationales. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1075–1088.
- Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. [Evaluating and characterizing human rationales](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [ELECTRA: Pre-training text encoders as discriminators rather than generators](#). In *ICLR*.
- 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.
- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.
- Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. 2024. Evaluating human alignment and model faithfulness of llm rationale. *arXiv preprint arXiv:2407.00219*.
- 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.

- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Mareike Hartmann and Daniel Sonntag. 2022. [A survey on improving NLP models with human explanations](#). In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 40–47, Dublin, Ireland. Association for Computational Linguistics.
- Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39.
- Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32.
- Terne Sasha Thorn Jakobsen, Laura Cabello, and Anders Sogaard. 2023. Being right for whose right reasons? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1033–1054.
- Dongjun Jang, Sungjoo Byun, and Hyopil Shin. 2024. [A study on how attention scores in the BERT model are aware of lexical categories in syntactic and semantic tasks on the GLUE benchmark](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1684–1689, Torino, Italia. ELRA and ICCL.
- 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.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). Preprint, arXiv:2310.06825.
- Sepehr Kamahi and Yadollah Yaghoobzadeh. 2024. [Counterfactuals as a means for evaluating faithfulness of attribution methods in autoregressive language models](#). In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 452–468, Miami, Florida, US. Association for Computational Linguistics.
- Jonathan Kamp, Lisa Beinborn, and Antske Fokkens. 2024. [The role of syntactic span preferences in post-hoc explanation disagreement](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16066–16078, Torino, Italia. ELRA and ICCL.
- Anastasia Kozlova, Albina Akhmetgareeva, Aigul Khanova, Semen Kudriavtsev, and Alena Fenogenova. 2024. Transformer attention vs human attention in anaphora resolution. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 109–122.
- Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. Post hoc explanations of language models can improve language models. *Advances in Neural Information Processing Systems*, 36:65468–65483.
- Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv e-prints*, pages arXiv–1907.
- Josephine Lukito, Bin Chen, Gina M. Masullo, and Natalie Jomini Stroud. 2024. [Comparing a BERT classifier and a GPT classifier for detecting connective language across multiple social media](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19140–19153, Miami, Florida, USA. Association for Computational Linguistics.
- Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024a. [Are self-explanations from large language models faithful?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.
- Andreas Madsen, Siva Reddy, and Sarath Chandar. 2024b. Faithfulness measurable masked language models. In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Supriya Manna and Niladri Sett. 2024. [Faithfulness and the notion of adversarial sensitivity in NLP explanations](#). In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 193–206, Miami, Florida, US. Association for Computational Linguistics.
- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

- Michael Neely, Stefan F Schouten, Maurits Bleeker, and Ana Lucic. 2022. A song of (dis) agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing. In *HHAI2022: Augmenting Human Intellect*, pages 60–78. IOS Press.
- Anabel Pilicita and Enrique Barra. 2025. Llms in education: Evaluation gpt and bert models in student comment classification. *Multimodal Technologies and Interaction*, 9(5):44.
- 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.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners.](#) *OpenAI Blog*, 1(8).
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks.](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Alexis Ross, Matthew Peters, and Ana Marasovic. 2022. [Does self-rationalization improve robustness to spurious correlations?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI’17*, page 2662–2670. AAAI Press.
- Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.
- 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.
- Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62.
- Zian Su, Xiangzhe Xu, Ziyang Huang, Zhuo Zhang, Yapeng Ye, Jianjun Huang, and Xiangyu Zhang. 2024. Codeart: Better code models by attention regularization when symbols are lacking. *Proceedings of the ACM on Software Engineering*, 1(FSE):562–585.
- Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. [Introduction to the CoNLL-2000 shared task: Chunking.](#) In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition.](#) In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9048–9056.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.](#) *Preprint*, arXiv:2412.13663.
- Sarah Wiegrefe and Ana Marasović. 2021. [Teach me to explain: A review of datasets for explainable nlp.](#) In *Proceedings of NeurIPS*.
- T Wolf. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Dean Wyatte, Fatemeh Tahmasbi, Ming Li, and Thomas Markovich. 2024. Scaling laws for discriminative classification in large language models. *arXiv preprint arXiv:2405.15765*.
- Sean Xie, Soroush Vosoughi, and Saeed Hassani-pour. 2024. [IvRA: A framework to enhance attention-based explanations for language models with interpretability-driven training.](#) In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 431–451.

Miami, Florida, US. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4094–4103.

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.

A Appendix

This appendix includes a set of secondary results (A.1), a report of the main technical details of our experiments (A.2) and a collection of complementary overview visualisations (A.3).

A.1 Secondary Results

BERT’s \mathcal{M} and \mathcal{R} are more stable than Pythia

By explicitly guiding the model to use rationales for a prediction, we expected that the agreement between models fine-tuned on different random seeds would stabilise. Through Fleiss’ κ (Fleiss, 1971), we measure the inter-model agreement between class label predictions on the test set, from three differently seeded runs. Figure 6 shows that BERT predictions are more stable both on the base model \mathcal{M} and on the model with attention regularisation \mathcal{R} . Besides the overall higher stability, BERT also exhibits more agreement increases (10 vs. 8) and the strongest improvement (+.09 on HateXplain_U). Pythia has the biggest drop (−.06 on CoNLL-chunk_{ADVP}).

CI_{rem} and CI_{msk} strongly correlate for BERT

Prior assumptions stated that different implementations of CI , i.e. by removing (CI_{rem}) or by masking (CI_{msk}) the non-rationales from the input, may yield contrasting results. Figure 7 shows that different implementations of CI are similar for BERT and Pythia. We compare the absolute Contextual Impact scores (correlation and polarity) in the light of a model comparison between BERT and Pythia. We expect the differences to be smaller for BERT as it should recognise masked and removed tokens in a similar way due to its masked language modelling in pretraining. We find that the correlation between the two is positive strong for both models (full

overview in Table 2. Only Kendall’s correlation is significant for BERT ($\tau = .87, p < .05$), which is however notable given the small sample. We find that 13 out of 14 task versions have same polarity for BERT, and 12 for Pythia. The slightly higher polarity error and the non-significant correlation by Pythia can be explained by the lack of an associated meaning of the [MASK] token for not being pre-trained as a masked language model. However, the difference is minimal, indicating that it may still be appropriate to compute CI through removal or masking interchangeably for left-to-right models.

Variables	BERT		Pythia	
	r	τ	r	τ
$CI_{\text{rem}}, CI_{\text{msk}}$.94	.87*	.83	.73
CI_{rem}, TC	.14	.07	-.09	.07
CI_{rem}, AR	-.49	-.33	.14	.20
CI_{msk}, TC	.09	-.07	.09	.07
CI_{msk}, AR	-.54	-.47	-.26	-.07

Table 2: Pairwise correlations for BERT and Pythia. The significance (*) threshold is set at $p < .05$.

CI diverges with \mathcal{R}

We compare the average, absolute distance from 0 for CI_{rem} and CI_{msk} . The differences between the two metrics are negligible for both models. Standard deviations are larger for BERT (.11 and .11 versus .05 and .04). Until this moment, we measured CI for the baseline model \mathcal{M} , but we measure it for the regularised model \mathcal{R} , too. If \mathcal{R} learnt from the rationales and CI is reflecting this, the difference in probabilities on the contrast examples (i.e. CI) should be closer to 0 than it was for \mathcal{M} . Table 3 shows that this is not the case: CI scores for \mathcal{R} are instead more distant and/or exhibit a greater standard deviation.

Polarity does not imply successful modelling

We explored whether negative CI relates to successful TC training (>1), which resulted to be true in a smaller percentage of the cases. Similarly, we find no pattern in the inverse relation between CI polarity and AR success, but this hypothesis is more likely to be true for Pythia (see Table 4).

BERT displays ‘positive’ behaviour

In Figure 2, we see that CI for BERT is mostly higher and more often positive than Pythia. This may indicate that, for BERT, uncontextualised rationales are less informative than for Pythia.

Learning from rationales entails CI thresholds

The results in Figure 4 and Figure 5 are reflected

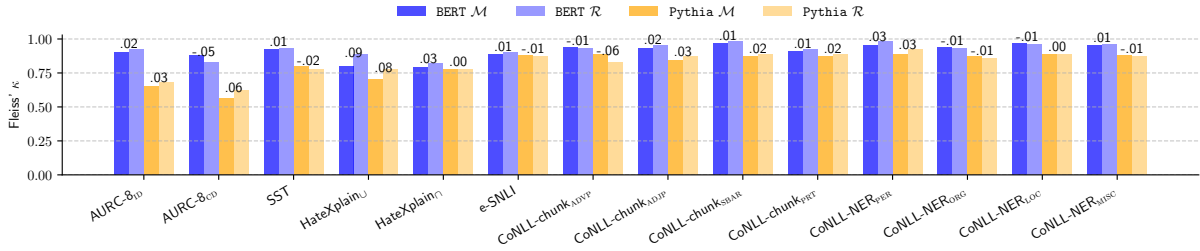


Figure 6: Inter-model agreement (Fleiss’ κ) on predictions given by three different runs. The agreement differences between base model \mathcal{M} and regularised model \mathcal{R} are reported on top of each of the model–task combinations.

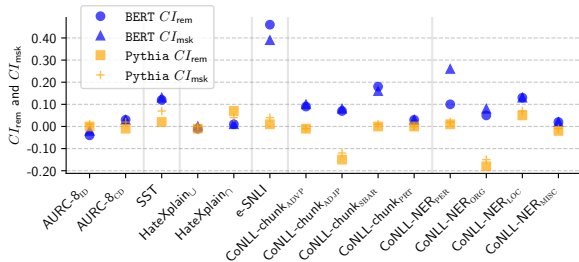


Figure 7: Different implementations of Contextual Impact compared: CI_{rem} and CI_{msk} are similar.

in Figure 8. Here, we see that for BERT–AURC-8_{cd} (the model that learned most from attention regularisation), for all models \mathcal{M} , \mathcal{R} and \mathcal{T} , sufficiency is mostly cluttered around 0, but that the extremes show a different behaviour. While there is no indication of correlation for \mathcal{T} , \mathcal{M} and \mathcal{R} show a clear threshold: below -0.50 the predictions are incorrect, above 0.50 the predictions are correct.

Model	BERT		Pythia	
	CI_{rem}	CI_{msk}	CI_{rem}	CI_{msk}
\mathcal{M}	.10 \pm .11	.10 \pm .11	.04 \pm .05	.04 \pm .04
\mathcal{R}	.10 \pm .11	.11 \pm .13	.05 \pm .08	.04 \pm .06

Table 3: Average sufficiency scores for \mathcal{M} and \mathcal{R} .

Metric	BERT		Pythia	
	CI_{rem}	CI_{msk}	CI_{rem}	CI_{msk}
TC	29%	21%	50%	36%
AR	29%	21%	36%	36%

Table 4: Tasks where polarity is a performance indicator.

A.2 Technical Details

Special tokens Sentences in e-SNLI are concatenated with `<sep>` token in BERT setup, but no separator token is used for Pythia (`eos_token` would not take into consideration the context left of the

token). `[MASK]`, `[UNK]` and `[PAD]` tokens are not present in Pythia tokenizer. Knowing that they are interpreted differently than in the BERT setup, but to avoid subtokenization (e.g. into `"["`, `"MASK"`, `"]"`), we add them to the pre-assigned placeholders in the Pythia tokenizer (`mask|unk|pad`)_token as random embeddings.

Rationale–token alignment Humans annotate rationales at the word level. When splitting words into (sub)tokens, we need to re-align the rationales. Before tokenising, we assign a rationale label at the character-level. After tokenising, we ignore the subtoken indicators `###` and `G` and remove non-ascii characters for which BPE tokenisation adds artifact characters to the tokenised strings, interfering with alignment. We then collapse character-level rationales to token-level rationales.

Software and hparams For each `<name>` among (Bert|GPTNeoX|ModernBert|GPTNeo) we use `<name>ForTokenClassification` and a custom `<name>Model` from Huggingface’s Transformers library, v4.44.2 (Wolf, 2019). We fine-tune for up to 10 epochs with a batch size of 16 and keep the checkpoint with the lowest evaluation loss. We kept the learning rate for sequence classification at $3e-5$ with some exceptions: $1e-5$ for GPTNeo and $3e-6$ for some of the ModernBERT runs. $3e-5$ was used for token classification. A learning rate decay of $1e-2$ was used in all experiments. More details are found in our repository.

Hardware and runtimes Table 5: models are fine-tuned on either NVIDIA A100-SXM4-40GB (40960 MiB) or on NVIDIA GeForce RTX 2080 Ti (11264 MiB), the choice of which does not affect runtime of the same setup with different random seeds (relatively small models and constant training batch sizes for comparability reasons). Rationalisation of the input does not affect runtime, nor does the dataset version. Token classifier runtimes are

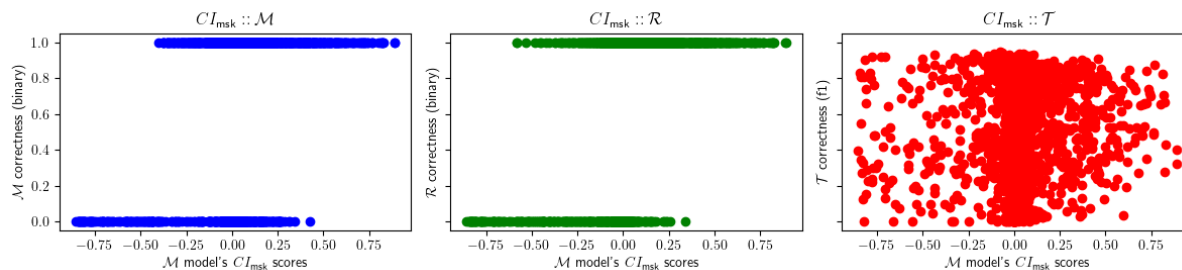


Figure 8: Sufficiency scores in relation to prediction correctness. In this example: CI_{msk} , BERT, AURC-8_{CD}.

comparable.

Dataset	Runtime
AURC-8	3–4 min
SST	4–6 min
HateXplain	17–25 min
e-SNLI	7–8 h
CoNLL-chunk	5–8 min
CoNLL-NER	13–19 min

Table 5: Approximate runtime of a single model fine-tuning on sequence classification tasks for 5 epochs.

Rationale derivation algorithm for SST We follow Carton et al. (2020) to derive the binary token-level rationale mask from the raw SST dataset. We report their description for completeness in the present appendix. Specifically, SST contains syntactic parse trees of movie reviews, where each node (i.e. phrase) is annotated with a sentiment label ranging from -2 (very negative) to +2 (very positive). Starting with the root and traversing the tree breadth-first, we include a node in the rationale (i.e. we assign a value of 1 to each of the node’s tokens) if the sentiment score of the node is greater than all of its constituents. In this way, only the smallest constituents that explain the sentiment are added to the rationale, maximising the granularity of the rationale.

A.3 Complementary Overview Visualisations

Figure 9 visualises the performance improvements of the classifiers along with their baselines. We then share the full learnability metrics’ scores in Table 6. Table 7, instead, represents the full overview of the confidence intervals for AR performance. Finally, Table 8 contains the computations of ΔPred and accompanies Figure 5 in §5.4.

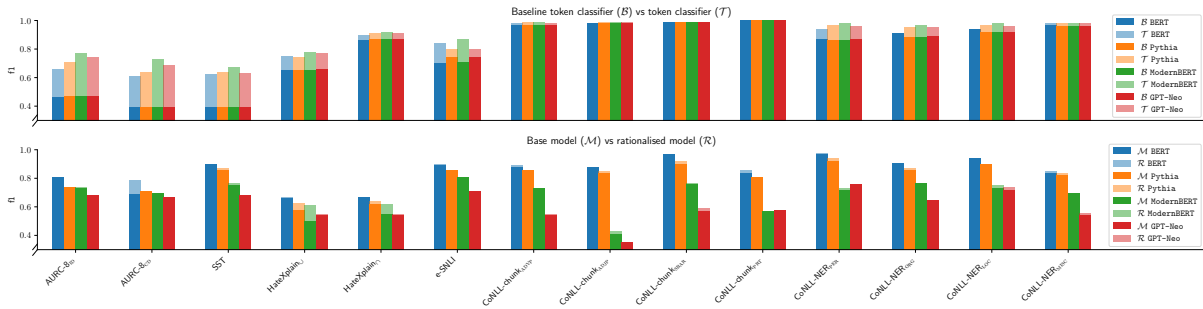


Figure 9: Performance improvements (f1) of the classifiers trained with rationales compared to the baselines.

Task versions	BERT		Pythia		ModernBERT		GPT-Neo	
	<i>TC</i>	<i>AR</i>	<i>TC</i>	<i>AR</i>	<i>TC</i>	<i>AR</i>	<i>TC</i>	<i>AR</i>
AURC-8 _{ID}	1.43 _{.463}	1.01 _{.808}	1.53 _{.467}	0.93 _{.741}	1.66 _{.467}	1.01 _{.733}	1.58 _{.467}	0.99 _{.677}
AURC-8 _{CD}	1.57 _{.391}	1.14 _{.692}	1.62 _{.393}	0.93 _{.712}	1.86 _{.393}	0.97 _{.702}	1.74 _{.393}	0.97 _{.674}
SST	1.60 _{.388}	1.00 _{.898}	1.65 _{.388}	1.00 _{.864}	1.73 _{.388}	1.03 _{.748}	1.61 _{.393}	1.00 _{.676}
HateXplain _∪	1.15 _{.648}	1.01 _{.660}	1.13 _{.652}	1.07 _{.582}	1.20 _{.652}	1.22 _{.499}	1.18 _{.655}	1.02 _{.542}
HateXplain _∩	1.04 _{.863}	0.99 _{.673}	1.05 _{.866}	1.03 _{.621}	1.06 _{.866}	1.14 _{.547}	1.05 _{.869}	1.02 _{.541}
e-SNLI	1.21 _{.698}	1.00 _{.894}	1.09 _{.736}	1.00 _{.860}	1.23 _{.707}	0.98 _{.807}	1.09 _{.737}	0.99 _{.706}
CoNLL-chunk _{ADVP}	1.01 _{.974}	1.00 _{.885}	1.01 _{.974}	0.98 _{.859}	1.02 _{.974}	0.95 _{.731}	1.01 _{.973}	1.02 _{.539}
CoNLL-chunk _{ADJP}	1.00 _{.981}	1.00 _{.883}	1.01 _{.981}	1.01 _{.840}	1.01 _{.981}	1.05 _{.412}	1.01 _{.982}	0.99 _{.352}
CoNLL-chunk _{SBAR}	1.01 _{.986}	1.01 _{.967}	1.01 _{.986}	1.03 _{.895}	1.01 _{.986}	1.02 _{.755}	1.01 _{.985}	1.04 _{.570}
CoNLL-chunk _{PRT}	1.00 _{.997}	1.03 _{.838}	1.00 _{.997}	1.00 _{.810}	1.00 _{.997}	0.99 _{.569}	1.00 _{.997}	0.95 _{.577}
CoNLL-NER _{PER}	1.09 _{.867}	1.01 _{.967}	1.12 _{.862}	1.01 _{.924}	1.14 _{.862}	1.01 _{.722}	1.11 _{.866}	1.00 _{.755}
CoNLL-NER _{ORG}	0.90 _{.907}	0.99 _{.911}	1.07 _{.882}	1.01 _{.860}	1.10 _{.882}	1.00 _{.766}	1.07 _{.886}	1.00 _{.652}
CoNLL-NER _{LOC}	0.94 _{.938}	1.00 _{.935}	1.05 _{.921}	1.00 _{.900}	1.07 _{.921}	1.02 _{.734}	1.04 _{.923}	1.03 _{.723}
CoNLL-NER _{MISC}	1.01 _{.968}	1.00 _{.844}	1.02 _{.962}	1.03 _{.817}	1.02 _{.962}	0.98 _{.702}	1.01 _{.964}	1.03 _{.539}

Table 6: Overview of learnability metrics. *TC* and *AR* are both ratios and equal an f1-score divided by a baseline; they are presented with their respective baselines, which are given in subscript. Example: the AURC-8_{ID} regularised BERT model \mathcal{R} outperforms the baseline model \mathcal{M} (f1 = .808) by 1.01 times (a 1% increase).

Task versions	BERT	Pythia	ModernBERT	GPT-Neo
AURC-8 _{ID}	1.01 [1.0022, 1.0090]	0.93 [0.8467, 0.9883]	1.01 [0.9926, 1.0213]	0.99 [0.9713, 1.0133]
AURC-8 _{CD}	1.14 [1.1035, 1.1759]	0.93 [0.8265, 1.0454]	0.97 [0.9339, 1.0167]	0.97 [0.9113, 1.0145]
SST	1.00 [0.9899, 1.0099]	1.00 [0.9881, 1.0195]	1.03 [0.9855, 1.0627]	1.00 [0.9626, 1.0396]
HateXplain _U	1.01 [0.9846, 1.0288]	1.07 [1.0034, 1.1437]	1.22 [1.1850, 1.2522]	1.02 [0.9944, 1.0494]
HateXplain _∩	0.99 [0.9724, 1.0154]	1.03 [0.9936, 1.0600]	1.14 [1.0410, 1.2270]	1.02 [0.9741, 1.0651]
e-SNLI	1.00 [0.9946, 1.0118]	1.00 [0.9981, 1.0055]	0.98 [0.9700, 0.9905]	0.99 [0.9900, 0.9962]
CoNLL-chunk _{ADVP}	1.00 [0.9974, 1.0123]	0.98 [0.9633, 0.9948]	0.95 [0.8635, 0.9949]	1.02 [0.9066, 1.1898]
CoNLL-chunk _{ADJP}	1.00 [0.9847, 1.0085]	1.01 [0.9967, 1.0239]	1.05 [0.8678, 1.2328]	0.99 [0.7409, 1.3584]
CoNLL-chunk _{SBAR}	1.01 [1.0041, 1.0112]	1.03 [1.0075, 1.0590]	1.02 [0.9872, 1.0539]	1.04 [0.9660, 1.1365]
CoNLL-chunk _{PRT}	1.03 [1.0071, 1.0514]	1.00 [0.9807, 1.0271]	0.99 [0.8954, 1.1073]	0.95 [0.8306, 1.0991]
CoNLL-NER _{PER}	1.01 [1.0008, 1.0303]	1.01 [0.9988, 1.0275]	1.01 [0.9948, 1.0343]	1.00 [0.9865, 1.0226]
CoNLL-NER _{ORG}	0.99 [0.9864, 0.9960]	1.01 [0.9913, 1.0251]	1.00 [1.0011, 1.0095]	1.00 [0.9661, 1.0477]
CoNLL-NER _{LOC}	1.00 [0.9969, 1.0034]	1.00 [0.9956, 1.0106]	1.02 [0.9512, 1.0783]	1.03 [0.9953, 1.0567]
CoNLL-NER _{MISC}	1.00 [0.9945, 1.0138]	1.03 [1.0055, 1.0548]	0.98 [0.9616, 1.0056]	1.03 [0.9995, 1.0650]

Table 7: Mean and 95% [confidence intervals] for AR performance across models and tasks.

Model	$\uparrow AR$			$\downarrow AR$		
	Top	Bottom	$\Delta Pred$	Top	Bottom	$\Delta Pred$
BERT	95	33	62	69	54	15
Pythia	87	12	75	84	46	38
ModernBERT	81	72	9	73	69	4
GPT-Neo	72	71	1	67	74	-7

Table 8: Accompanying table for Figure 5. Comparison of top AR (\uparrow) and bottom AR (\downarrow) correct predictions for the rationalised model \mathcal{R} . We compare their signed differences denoted as $\Delta Pred$.