# Human Rationales as Attribution Priors for Explainable Stance Detection

**Sahil Jayaram**
Columbia University
New York, NY
sahil.j@columbia.edu

**Emily Allaway**
Columbia University
New York, NY
eallaway@cs.columbia.edu

## Abstract

As NLP systems become better at detecting opinions and beliefs from text, it is important to ensure not only that models are accurate but also that they arrive at their predictions in ways that align with human reasoning. In this work, we present a method for imparting human-like rationalization to a stance detection model using crowdsourced annotations on a small fraction of the training data. We show that in a data-scarce setting, our approach can improve the reasoning of a state-of-the-art classifier—particularly for inputs containing challenging phenomena such as sarcasm—at no cost in predictive performance. Furthermore, we demonstrate that attention weights surpass a leading attribution method in providing faithful explanations of our model's predictions, thus serving as a computationally cheap and reliable source of attributions for our model.

## 1 Introduction

Stance detection, automatically identifying the position on a topic taken by a text (Mohammad et al., 2017), allows readers to glean valuable information from news articles and social media, such as whether the writing is politically slanted. Due to the sensitive nature of many topics (e.g., political ideologies and religious beliefs), it is crucial that stance models are transparent and rationalize their predictions in human-like ways. Furthermore, their reasoning must *remain* human-like even when they are tasked with generalizing to a new, unseen test topic.

A model's rationale for a specific input can be extracted in the form of *feature attributions*, which quantify the influence that each input feature asserts on the model's predictions. Methods for obtaining these attributions vary by *faithfulness* (the extent to which they accurately measure the importance of each feature). Attributions can be incorporated into the training process via *attribution priors*, a powerful framework for imparting domain knowledge to models (Erion et al., 2019, 2020). In this work, we propose a prior that penalizes a stance classifier for producing attributions that deviate from human judgements of word importance (i.e., which words or phrases in a text are most indicative of stance). Notably, our method is model-agnostic and can be used with any differentiable feature attribution technique.

| Topic = *marijuana* | True Label = *FOR* |
|---|---|
| **Baseline** | The better question is, though, how many people who have used marijuana DO NOT use other illegal drugs. I think the answer would surprise you. Correlation is not causation - the sooner you learn that, the sooner you can get to the real root of our opiate addition problem. |
| Prediction | AGAINST (wrong) |
| **Our Method** | The better question is, though, how many people who have used marijuana DO NOT use other illegal drugs. I think the answer would surprise you. Correlation is not causation - the sooner you learn that, the sooner you can get to the real root of our opiate addition problem. |
| Prediction | FOR (correct) |

Table 1: Model reasoning as explained by mean attention weights (MAW). The baseline is trained using only cross-entropy loss, while the second model is trained using our proposed attribution prior.

We train and evaluate our model on the VAST (VAried Stance Topics) dataset, whose test set covers topics that are either absent from (zero-shot) or scarce in (few-shot) the training data (Allaway and McKeown, 2020). To construct our attribution prior, we crowdsource word importance annotations on a small subset ($\sim$500 examples) of the training data, as well as on a sample of the test



*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5540–5554
November 7–11, 2021. 

data for evaluation. Additionally, to assess how our method might fare in realistic, resource-limited scenarios, we experiment using not only the complete VAST train set, but also reduced versions of it.

As the attribution method for our prior, we choose mean attention weights (MAW), a method that is computationally cheap in comparison to popular alternatives. Although numerous recent publications have demonstrated that attention weights do not *in general* provide faithful explanations of model behavior, we find strong evidence that the attributions offered by MAW are more faithful to our model's predictions than Gradient $\times$ Input (GI), a leading attribution method for transformer models.

Our contributions are as follows: (1) we propose a method that, in a simulated data-scarce setting, improves the reasoning of a state-of-the-art stance detection model without compromising its performance and (2) we show that MAW is not only simpler and computationally cheaper than GI, but also more faithful to our models. Our data and models are available at https://github.com/SahilJ97/Explainable-Stance-Detection.

## 2 Related Work

While early work on stance detection focused primarily on ideological debates (Walker et al., 2012; Hasan and Ng, 2014; Abbott et al., 2016), recent datasets have also begun to include more political topics, such as elections (Mohammad et al., 2016; Vamvas and Sennrich, 2020; Lai et al., 2020) and referendums (Taulé et al., 2017; Tsakalidis et al., 2018). This reflects a growing interest in developing models to understand public opinion on a range of topics. However, to be used in real-world scenarios, such models must also exhibit good generalization ability and a degree of transparency. Recent work has focused on the generalization ability of stance detection models: across topics (Augenstein et al., 2016; Xu et al., 2018; Allaway and McKeown, 2020; Zhang et al., 2020; Allaway et al., 2021), languages (Vamvas and Sennrich, 2020), and even label sets and genres (Schiller et al., 2020; Hardalov et al., 2021). In contrast, our work focuses on the *reasoning* of models during topic generalization.

Many attribution methods have been used to extract rationales from text classifiers, including activation-based (Atanasova et al., 2020), perturbation-based (Ribeiro et al., 2016), gradient-based, and attention-based (Abnar and Zuidema, 2020; Wu and Ong, 2021) methods. Furthermore, numerous works incorporate these model attributions into the training process. Liu and Avci (2019) train a toxicity classifier using an attribution prior based on Integrated Gradients, a gradient-based method of feature attribution (Sundararajan et al., 2017). Zhong et al. (2019) directly train an attention mechanism for relation extraction. Previous studies have used word importance annotations much like ours to supervise attention with the goal of improving predictive performance (Pruthi et al., 2020; Kanchinadam et al., 2020). Our work, in contrast, focuses on verifiably improving the *reasoning* of a classifier. While there exist models whose reasoning paths are naturally transparent and easy to train, such as *select-predict* (Jacovi and Goldberg, 2021) and *rationale-augmented* (Zaidan et al., 2007; Zhang et al., 2016) models, our framework makes no assumptions about model architecture, and can thus be applied to a state-of-the-art stance classifier.

A recent survey of explainability (i.e., attribution) methods proposed five diagnostic properties for comparing techniques, including *faithfulness*, defined as a measure of how true attributions are to the inner workings of a model (Atanasova et al., 2020). Their experiments with transformer models show that gradient-based methods (e.g., GI) score high in faithfulness. In fact, similar assessments of faithfulness have found that attention does not provide faithful explanations (Jain and Wallace, 2019; DeYoung et al., 2020), especially compared to GI (Wu and Ong, 2021). However, recent studies have argued for a more nuanced understanding of faithfulness that prioritizes 'explainable enough' (Wiegreffe and Pinter, 2019; Jacovi and Goldberg, 2020) and allows for the 'best' technique to vary by model, task, or input (Ghorbani et al., 2018). Our work presents a scenario in which attention is in fact more faithful than GI.

## 3 Data

### 3.1 Crowdsourcing Annotation

In order to study model rationales for stance detection across many topics, we annotate a portion of the recently proposed VAST (Allaway and McKeown, 2020) dataset with *human* rationales. VAST is composed of comments (referred to here as *arguments*) from a portion of The New York Times. The stance topics in the dataset were extracted automat-

ically (e.g., by identifying important noun-phrases) and then validated (or corrected) using crowdsourcing. Crowdsourcing was also used to assign a label to each example: "pro" (for), "con" (against), or "neutral."

For annotation, we randomly select from the train set 700 non-neutral examples whose topics were validated by annotators. We also select 75 such examples from the test set. We do not annotate examples (142 in training, 32 in test) for which we judge the topic to be unclear (e.g., "problem"). Furthermore, in the training sample, we identify 14 examples with incorrect stance labels.

We collect our annotations via Amazon Mechanical Turk (MTurk), a popular crowdsourcing platform. For each HIT (task), workers are asked to (1) classify the stance of an argument with respect to a topic and (2) select the $k$ most *important* words in the argument (for each example, we provide an acceptable range of values for $k$). A word is considered to be *important* if masking it would make (1) more difficult[1]. A detailed illustration of our MTurk HIT is provided in Appendix A.

To ensure the quality of our annotations, we first publish a "qualification" HIT consisting of a single example. Then, three qualified workers annotate each example in our subsets. We also use (1) above to check for worker quality in the training subset only. In particular, for 74 samples in the training subset, at least two annotators disagree with the gold stance label. The authors inspect each of these examples and either flip the label (35 examples) or discard the example. The Cohen's $\kappa$ on these decisions is 0.392.

### 3.2 Oracle Attributions

**Computation and Processing**: The results of our crowdsourced tasks are used to compute *oracle attributions* for each of the annotated arguments.

We disregard a worker's annotations for an example if their stance classification (see (1) above) disagrees with the label. Our oracle attributions are a weighted sum of annotator responses, with worker quality score (WQS) as the weighting factor. WQS measures an annotator's word-level agreement with their peers (Dumitrache et al., 2018). The average WQS for our annotators is 0.58. Note that these oracle scores are later normalized (§4.2).

[1]The actual definition provided in our HITs differs somewhat (see Appendix A).

**Analysis**: We examine the processed oracle attributions and observe that annotators mark an average of 26% of the tokens in an argument as important. Surprisingly, words from the topic are only selected for 51% (267/519) of the examples while an average of 44% of important words are stopwords. For example, a worker selected "no one here" in the sentence "I know of no one here who is even remotely excited about the olympics" (see Table 5). This shows that human word importance judgements cannot be approximated simply by selecting the topic or the words most similar to the topic. Additionally, we find that on average only 10% (9%) of important words are positive (negative) sentiment-bearing, as identified by the MPQA lexicon (Wilson et al., 2017). This further highlights the complexity of human word importance judgements for stance detection, since sentiment has only a minor role in determining stance.

## 4 Methods

We propose a model for stance detection that uses a BERT-based encoder (§4.1) trained with an additional loss term (§4.2) designed to impose a prior based on human rationales.

Define $D = \{x_i = (d_i, t_i, y_i)\}_{i=1}^{N}$ as a dataset with $N$ examples, each consisting of an argument $d_i$, a topic $t_i$, and a stance label $y_i \in \{0, 1, -1\}$. In addition, let $M_\theta$ be some model with parameters $\theta$. Then we can define for each example $x$ a set of oracle attributions $s$, model attributions $a$, and *penalty weights* $\gamma$ (the contribution of each token to our loss term). Our *prior loss* term encourages the model to produce, for each example, attributions $a$ that are very "similar" to $s$ (§4.2-4.3).

### 4.1 Base Architecture

Our base architecture builds on the baseline model `BERT-joint` introduced by Allaway and McKeown (2020), which jointly embeds a topic and document using BERT (Devlin et al., 2019), thereby conditioning the topic representation on the document and vise versa. Our model differs from `BERT-joint` in two ways. First, rather than fixing BERT, we fine-tune its weights during training, thus allowing the transformer to update its attention heads. This is necessary in order to accommodate our choice of attribution method (MAW). Second, rather than removing stopwords from the input, we

pass to BERT the full input sequence ([CLS] document [SEP] topic [SEP]) and compute the final representation used for stance classification by taking the mean hidden state over all non-stopwords. We do this because our oracle attributions cover all words in the argument, not just the non-stopwords.

### 4.2 Prior Loss

Our example-level *rationale loss* function $\Omega$ is a weighted mean square error between normalized model attributions and normalized oracle attributions. Formally, for example $x = (d, t, y)$, let $m$ be the length of our argument $d$ and let $\theta$ denote the parameters of our model. Let $x$ be a word-importance-annotated example with penalty weights $\gamma = (\gamma_1, ..., \gamma_m)$, oracle attributions $s = (s_1, ..., s_m)$, and model attributions $a = (a_1, ..., a_m)$. Let $a'_j$ denote the *normalized* attribution score for argument token $j$. That is,

$$a'_j = \frac{a_j}{\sum_{i=1}^{m} a_i}.$$

Similarly, let $s'_j$ denote the normalized oracle score for argument token $j$. Then $\Omega$ is defined as follows:

$$\Omega(\theta; x) = \frac{\sum_{j=1}^{m} \gamma_j (a'_j - s'_j)^2}{\sum_{j=1}^{m} \gamma_j}. \quad (1)$$

Intuitively, the square error associated with the attribution on the argument's $j$-th token is weighted by penalty weight $\gamma_j$. For an example $x'$ that is not endowed with oracle attributions, we define $\Omega(\theta; x')$ to be 0.

Our complete loss function is the sum of the stance classification loss $\mathcal{L}_c$ and the scaled average prior loss across examples

$$\mathcal{L}(\theta; D) = \mathcal{L}_c(\theta; D) + \lambda \mathcal{L}_p(\theta; D) \quad (2)$$

$$\mathcal{L}_p(\theta; D) = \frac{1}{N} \sum_{i=1}^{N} \Omega(\theta; x_i) \quad (3)$$

where $D$ is a dataset, $\mathcal{L}_c$ is the cross-entropy loss, and $\lambda > 0$ is a hyperparameter.

### 4.3 Penalty Weights

In order to exclude certain tokens (punctuation and numerals) from our rationale loss function, and to potentially assign non-uniform influence to the remaining tokens, we introduce the notion of *penalty weights*. The penalty weight of a token specifies its contribution to the rationale loss function. In our experiments, we focus primarily on *binary* penalty weights, where tokens that are punctuation marks or numerals receive a score of 0 and all other tokens receive a score of 1. However, we also experiment with *tf-idf* penalty weights, where the score of a non-punctuation, non-numeral token is its *tf-idf* with respect to the training set.[2]

### 4.4 Feature Attribution

Our prior allows for any choice of attribution method. We select Mean Attention Weights (MAW) because of its extremely low computational cost in comparison with other methods, most of which require backpropagation and/or multiple forward passes (Atanasova et al., 2020). In MAW, the attribution score of token $j$ is the mean, taken across all tokens, layers, and attention heads, of attention weights $\alpha_{ij}$—that is, all attention weights associated with the key at index $j$. Informally, MAW measures how much attention each token *receives* (from other tokens as well as from itself). Our framework assumes that attribution scores are magnitudes (i.e., unsigned). Thus, we implicitly take the absolute value of our MAW attributions.

We also compare MAW with an additional attribution method, Gradient $\times$ Input (GI) (Wu and Ong, 2021), for evaluation. Let $e_j = (e_{j1}, \ldots, e_{jh})$ be the input embedding of the $j$th token of the argument for some example $x$. We then define the GI attribution score of token $j$ as

$$a_j^{GI} = \sqrt{\sum_c \sum_{k=1}^{h} \left( \frac{\partial f_c(x)}{\partial e_{jk}} e_{jk} \right)^2} \quad (4)$$

where $f_c$ denotes the component of the model's output function corresponding to class $c$. Intuitively, GI measures the sensitivity of the model to perturbations of $e_j$, which in theory measures the dependence of the model's prediction on token $j$. We choose to aggregate across output classes because all output neurons contribute to the model's decision; a negative contribution to a non-predicted class is just as important as a positive contribution of equal magnitude to the predicted class (Bach et al., 2015). We select GI as a benchmark method because of its high performance in assessments of faithfulness for transformer-based models across several domains (Atanasova et al., 2020; Wu and Ong, 2021).

[2] An example's "document" is the union of all arguments in the dataset that are associated with the example's topic.

| | Num. Ex | % with Oracle | Num. Topics$_f$ | Num. Topics$_z$ |
|---|---|---|---|---|
| **full** | 13438 | 3.9 | 5019 | - |
| **reduced**$_{25}$ | 3801 | 13.7 | 1831 | - |
| **reduced**$_{10}$ | 1788 | 29.0 | 887 | - |
| **Dev** | 2062 | - | 114 | 383 |
| **Test** | 3006 | 1.4 | 159 | 600 |

Table 2: Dataset statistics for VAST and training settings. $f$ indicates few-shot topics, $z$ indicates zero-shot topics.

## 5 Experiments

### 5.1 Data

We train and evaluate our models on VAST, using the standard train/dev/test split and both subsets of the test set: *few-shot* (few training examples per test topic) and *zero-shot* (no training or development examples per test topic). Bearing in mind that obtaining high-quality stance data across a broad range of topics is extremely resource-intensive (Allaway and McKeown, 2020), we assess the practical value of our approach by evaluating it under different degrees of data scarcity. Specifically, we experiment using three data settings that vary the number of training examples *without* oracle attributions: in addition to using all training examples (**full**), we also experiment using only a random sample of 25% (**reduced**$_{25}$) or 10% (**reduced**$_{10}$) of training examples *without* oracle attributions. We use all 519 of the training examples with oracle attributions in all three data settings (see Table 2).

### 5.2 Models

We train a stance model (**prior-bin:gold**) with our proposed attribution prior, using binary penalty weights (§4.3), our crowdsourced oracle attributions (§3.2), and MAW to extract model attributions (§4.4). We compare this model to one that shares its architecture but is trained without prior loss (**base**). In addition, we compare with two baselines proposed for VAST: **BERT-joint** – our architecture (§4.1) without fine-tuning and with additional data pre-processing, and **TGA Net** – a modification of `BERT-joint` that uses unsupervised clustering and attention to improve performance on unseen topics (Allaway and McKeown, 2020).

We tune $\lambda$ using a manual hyperparameter search. We find that because only a small fraction of examples are endowed with oracle attributions, the coefficient applied to our prior loss term must be quite large: $\lambda = 49152$ in the `full` and `reduced`$_{25}$ settings and $\lambda = 16384$ in the `reduced`$_{10}$ setting. Our models are implemented in PyTorch[3] and optimized using Adam for 20 epochs with a batch size of 32 and a fixed learning rate of $10^{-5}$. We use a maximum sequence length of 250 for arguments and 10 for topics. All models use bert-base-uncased from Huggingface[4]. Results are averaged across three random seeds unless otherwise specified.

### 5.3 Results: Stance Prediction

We evaluate our models using macro-averaged F1 on both the few-shot and zero-shot subsets (§3.1) of the VAST test set (see Table 3). We see that across training settings, `prior-bin:gold` and `base` achieve comparable results and outperform the baselines proposed for VAST. We also conduct an ablation on the method for computing penalty weights in the prior loss (§4.3) in the data-scarce reduced$_{25}$ setting. Specifically, we experiment with **prior-tfidf:gold** – *tf-idf* penalty weights and crowdsourced oracle attributions and **prior-bin:tfidf** – binary penalty weights and *tf-idf* values as pseudo-oracle attributions (instead of our crowdsourced labels). Both these methods perform worse than `prior-bin:gold` and `base`, achieving 0.661 and 0.655 macro-F1 respectively. This result aligns with our observations about human word importance annotations (§3.2), namely that human rationales are complex and do not necessarily parallel notions of word importance derived through *tf-idf*. Therefore, our stance prediction results show that human word importance annotations are necessary in order to obtain strong results using our proposed attribution prior.

### 5.4 Analysis of Rationales

In addition to evaluating our models' predictions, we also assess the quality of their reasoning. In order to do this, we first analyze the relative reliability of explanations obtained from MAW and GI. We then use our findings to evaluate rationale quality via two separate mechanisms: human raters and our rationale loss function ($\Omega$).

**Faithfulness of Attributions**: The *faithfulness* of an attribution method is the extent to which it accurately reflects a model's reasoning (Herman,

[3] https://pytorch.org
[4] https://huggingface.co/transformers

| | | All | | | Zero-Shot | | | Few-Shot | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Pro | Con | Avg | Pro | Con | Avg | Pro | Con | Avg |
| full | BERT-joint† | .545 | **.591** | .653 | .546 | .584 | .661 | .544 | **.597** | .646 |
| | TGA Net† | .573 | .590 | .665 | .554 | **.585** | .666 | .589 | .595 | .663 |
| | base | .643 | .581 | **.692** | .632 | .563 | .692 | **.652** | **.597** | **.691** |
| | prior-bin:gold | **.645** | .546 | .684 | **.649** | .542 | **.693** | .641 | .549 | .669 |
| reduced$_{25}$ | BERT-joint | .516 | .524 | .603 | .553 | .527 | .619 | .480 | .522 | .587 |
| | base | .626 | **.559** | **.673** | .634 | **.564** | .688 | .618 | **.552** | **.658** |
| | prior-bin:gold | **.637** | .549 | **.673** | **.643** | .537 | **.694** | **.631** | .527 | .653 |
| reduced$_{10}$ | BERT-joint | .450 | .469 | .370 | .491 | .478 | .372 | .422 | .448 | .366 |
| | base | **.594** | .491 | .623 | **.600** | .460 | .630 | **.589** | .513 | **.614** |
| | prior-bin:gold | .579 | **.526** | **.630** | .596 | **.522** | **.650** | .562 | **.529** | .609 |

Table 3: F1 results on the test set for all three versions of the train set. Avg refers to the macro-average across all three classes (Pro, Con, and Neutral). † marks results reported in Allaway and McKeown (2020). Differences between base and prior-bin:gold are not statistically significant ($p < .05$).

| Method | prior-bin:gold | base |
|---|---|---|
| random | .353 | .358 |
| GI | .285 | .326 |
| MAW | **.264** | **.299** |

Table 4: Area under the threshold-performance curve (AUC-TP).

2017). Although Atanasova et al. (2020) propose five diagnostic properties for explainability techniques, we only consider faithfulness, as we find that the other four properties are either non-meaningful or inapplicable in the case of our methods (see Appendix B).

Our faithfulness analysis considers only the reduced$_{25}$ setting, as we are interested in improvements under data-scarcity and believe the faithfulness of MAW and GI to be relatively constant across all three data settings. To gauge the faithfulness of a feature attribution method, we use the diagnostic employed by Atanasova et al. (2020). Namely, for all $\psi \in \{0, 10, \ldots, 100\}$, we mask the most important (as determined by the attribution method) $\psi\%$ of tokens in each input example and compute the resulting macro-F1 across all examples. The area under this *threshold-performance curve* (AUC-TP) gives us an *inverse* measure of faithfulness; intuitively, if an attribution method is faithful, then model performance relies predominantly on the most important tokens as suggested by that method, resulting in a *low* AUC-TP. As a baseline, we also compute a threshold-performance curve using random masking (equivalent to assessing random attributions).

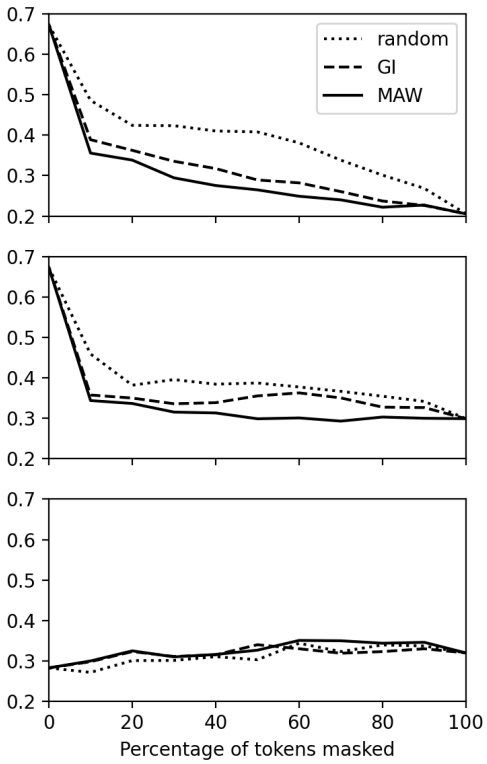


Figure 1: Threshold-performance curves for prior-bin:gold (top), base (middle), and an untuned model (bottom). Lower AUC-TP relative to random suggests more faithful attributions.

We find that MAW surpasses GI in faithfulness for both prior-bin:gold and base and considerably outperforms random attributions (see Figure 1, Table 4). This indicates that overall, MAW attributions are faithful to our model, and can therefore be trusted, for our purposes, as explanations of model reasoning. In other words, we can justifiably interpret MAW attributions as rationales.

| | Eval Scores | Attributions |
|---|---|---|
| oracle | | I have lived in brazil for the last five years (and off and on over the last 27 years). I know of no one here who is even remotely excited about the olympics. It would seem that people don't care. The economy is tanking and government is at a complete standstill. We have more important things on our mind right now. |
| `prior-bin: gold` | congruity: 2<br>sufficiency: 2<br>irrelevance: 0 | I have lived in brazil for the last five years (and off and on over the last 27 years). I know of no one here who is even remotely excited about the olympics. It would seem that people don't care. The economy is tanking and government is at a complete standstill. We have more important things on our mind right now. |
| `base` | congruity: 0<br>sufficiency: 0<br>irrelevance: 2 | I have lived in brazil for the last five years (and off and on over the last 27 years). I know of no one here who is even remotely excited about the olympics. It would seem that people don't care. The economy is tanking and government is at a complete standstill. We have more important things on our mind right now. |

Table 5: Human quality judgements of model rationales and comparison with oracle attributions.

**Human Evaluation of Rationales**: For both `prior-bin:gold` and `base`, we select the random seed with the lowest prior loss $\mathcal{L}_p$ on the test set. We sample 40 test set examples for which both models predict the correct stance and ask NLP researchers to rate the models' MAW attributions on:

1. **congruity**: how well the explanation matches the stance label
2. **sufficiency**: whether enough words are selected[5] that a human could infer the stance
3. **irrelevance**: how often irrelevant words are selected.

Ratings are done on a five-point Likert scale.

Each example is scored by two annotators. The Krippendorff's alpha (Krippendorff, 1980) for congruity, sufficiency, and irrelevance are 0.316, 0.311, and 0.136, respectively. Score averages are mapped to a 0-2 scale for analysis. As in our faithfulness analysis, we focus exclusively on the reduced$_{25}$ setting.

We find that `prior-bin:gold` outperforms `base` by a considerable margin for all three questions (see Table 6), demonstrating that in a data-scarce setting, our attribution prior is highly effective for conditioning model reasoning.

Additionally, we investigate whether rationale loss is a reliable proxy for rationale quality. Specifically, we compute the correlation between root rationale loss ($\sqrt{\Omega}$) and the averaged human evaluation scores. We use root loss since our rationale loss $\Omega$ is intuitively a weighted mean-square error (§4.2). We compute correlation for both

[5]Refer to Appendix C for information regarding how we generate visualizations of attributions.

| | `prior-bin:gold` | `base` |
|---|---|---|
| congruity↑ | **1.12** | 0.18 |
| sufficiency↑ | **1.18** | 0.18 |
| irrelevance↓ | **1.06** | 1.71 |

Table 6: Average human evaluation scores (0-2 scale) for MAW attributions. ↑ indicates that a higher score is preferred (↓, lower).

| | $\sqrt{\Omega}_{MAW}$ | $\sqrt{\Omega}_{GI}$ |
|---|---|---|
| congruity↑ | 0.241 | **0.317** |
| sufficiency↑ | 0.296 | **0.328** |
| irrelevance↓ | **-0.310** | -0.175 |

Table 7: Pearson correlation coefficients between human evaluation results for MAW attributions and root rationale loss ($\sqrt{\Omega}$), with the latter shown for both GI and MAW attributions. ↑ indicates that a positive correlation is preferred (↓, negative).

MAW-based ($\sqrt{\Omega}_{MAW}$) and GI-based ($\sqrt{\Omega}_{GI}$) root rationale loss for our models. We find that both metrics correlate with human judgements; however, while $\sqrt{\Omega}_{MAW}$ is a better indicator of irrelevance, $\sqrt{\Omega}_{GI}$ is a better indicator of congruity and sufficiency (see Table 7). This suggests that in the context of attribution *priors*, different attribution methods are better suited for enforcing different qualities on model rationales, and attribution methods should be chosen accordingly.

**Computational Evaluation of Rationales**: In terms of MAW-based rationale loss ($\Omega_{MAW}$), training with our attribution prior yields a statistically significant advantage in all data settings (see Table 8). In terms of $\Omega_{GI}$, the model trained with our proposed prior performs best in the moderately

| | | All | | Zero-Shot | | Few-Shot | |
|---|---|---|---|---|---|---|---|
| | | MAW | GI | MAW | GI | MAW | GI |
| full | base | .112 | **.121** | .833 | **.910** | .972 | **1.038** |
| | prior-bin:gold | **.107*** | .122 | **.805*** | .948 | **.924*** | 1.067 |
| reduced$_{25}$ | base | .112 | .122 | .827 | .894 | .967 | 1.058 |
| | prior-bin:gold | **.106*** | **.120** | **.803*** | **.876*** | **.919*** | **1.051** |
| reduced$_{10}$ | base | .113 | **.119** | .831 | **.877** | .977 | 1.042 |
| | prior-bin:gold | **.120*** | .120 | **.805*** | .917 | **.933*** | **1.039** |

Table 8: $\sqrt{\lambda \mathcal{L}_p}$ for MAW and GI on the VAST test set, with $\lambda = 16384$ for all settings. * indicates statistical significance ($p < .05$). Prior loss was computed over the entire test set for All and over the word importance-annotated subset for Zero-Shot and Few-Shot.

| | | *Imp* | | *mlT* | | *mlS* | | *Qte* | | *Sarc* | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MAW | GI | MAW | GI | MAW | GI | MAW | GI | MAW | GI |
| base | I | .890 | .820 | 1.187 | 1.049 | .938 | .841 | .818 | .758 | .909 | .831 |
| | O | .927 | .854 | .890 | .825 | .889 | .852 | .966 | .888 | **.961** | .921 |
| prior-bin:gold | I | **.869** | **.789** | **1.177** | **1.035** | **.917** | **.807** | **.800** | **.732** | **.888** | **.793** |
| | O | **.914** | **.811** | **.874** | **.782** | **.882** | **.804** | **.952** | **.841** | .976 | **.868** |

Table 9: $\sqrt{\lambda \mathcal{L}_p}$ for challenging phenomena in the reduced$_{25}$ data setting. I indicates examples containing the phenomenon and O indicates examples without it. All results are statistically significant ($p < .05$). Prior loss was computed as in Table 8.

data-scarce setting, but falls short of base when the complete train set is used. This may indicate that in the full setting, prior-bin:gold's rationales are poorer than those of base in terms of congruity and sufficiency. Thus, our attribution prior may have adverse effects on model reasoning when an insufficient fraction of the train set is endowed with oracle attributions.

### 5.5 Error Analysis

**Challenging Phenomena**: We also examine performance on the five challenging phenomena identified in VAST: *Imp* – the topic phrase is absent from the argument and the label is non-neutral, *mlT* – the argument appears in multiple examples (each with a different topic), *mlS* – the argument appears in multiple examples with different, non-neutral stance labels, *Qte* – the argument contains a quotation, and *Sarc* – the argument contains sarcasm. We find that while training with our proposed attribution prior yields comparable performance on these phenomena, it provides superior rationales for all five (see Table 9). This shows the efficacy of our method at improving rationalization for difficult examples without degrading performance.

**Rationale Error Types**: We analyze the errors in the *rationales* produced by prior-bin:gold. Specifically, we randomly sample 50 examples for which the model predicts the incorrect label and manually categorize them as: *amount*$_{err}$ – errors in the amount of words selected (i.e., selecting too few or too many), *content*$_{err}$ – errors in the content of selected words (i.e., missing negations or critical parts of phrases), *complex*$_{err}$ – failure to understand complex language (e.g., sarcasm or implicit references to the topic), and *data*$_{err}$ – errors in the data annotation (i.e., incorrect label or nonsensical topic). Semantic errors (content$_{err}$ and complex$_{err}$) occur in 68% of the cases (32% and 36% respectively). For example, the model often fails to understand rhetorical questions or misses important negations. Additionally, we find that 46% of the rationales select too few or too many words (e.g., selecting most stopwords in the argument). Finally, we see that data$_{err}$ account for 30% of the errors. This analysis suggests that, while our attribution prior improves the rationales for semantically complex examples, semantic understanding remains a key challenge for future improvements.

## 6 Conclusion

This paper addresses two issues concerning the task of stance detection: 1) the need for models whose reasoning aligns with that of humans and 2) the need for a way to meaningfully observe the reasoning of models in the first place. We find that

in a simulated data-scarce setting, our attribution prior improves model rationales using a practical volume of crowdsourced annotations. We also find that attention-based explanations, which have recently been the subject of much criticism, provide faithful explanations of our models' behavior, more so than a high-ranking alternative method.

In future work we plan to apply our method to more challenging settings, such as multilingual zero-shot stance detection. We will also further investigate the "economics" of our method—for instance, the number of annotated examples necessary to meaningfully improve model reasoning—as well as experiment with a broader range of attribution methods, e.g., Guided Backpropagation (Springenberg et al., 2015) and LIME (Ribeiro et al., 2016). Lastly, we hope to study how to condition model reasoning to protect against adversarial attacks.

## Acknowledgements

We thank Kathleen McKeown, the Columbia NLP group, and the anonymous reviewers for their comments. This work is supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1644869. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of the U.S. Government, or the NSF.

## 7 Ethics Statement

We use a dataset collected and distributed by Allaway and McKeown (2020): `https://github.com/emilyallaway/zero-shot-stance`. Data was collected from publicly available comments on articles on *The New York Times*. No user information is retained with the comments, so the data does not contain explicit information about race, gender or ethnicity of the original authors. For the additional annotations we collect, we compensate workers at ∼$13 per hour, above the federal minimum wage in the United States (where many annotators are based).

Some of the methods we discuss are intended to provide model transparency when predicting stance labels, including on sensitive topics (e.g., religious beliefs). When using these methods to provide explanations for a prediction on a text, real-world users should be informed that the explanations are automatically generated and may not be representative of the full opinions of the text's authors.

## References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn A. Walker. 2016. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *LREC*.

Samira Abnar and W. Zuidema. 2020. Quantifying attention flow in transformers. *ArXiv*, abs/2005.00928.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *EMNLP*.

Emily Allaway, Malavika Srikanth, and K. McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. *ArXiv*, abs/2105.06603.

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

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *EMNLP*.

S. Bach, Alexander Binder, Grégoire Montavon, F. Klauschen, K. Müller, and W. Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

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

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. *CoRR*, abs/1808.06080.

G. Erion, J. Janizek, Pascal Sturmfels, Scott M. Lundberg, and Su-In Lee. 2019. Learning explainable models using attribution priors. *ArXiv*, abs/1906.10670.

G. Erion, J. Janizek, Pascal Sturmfels, Scott M. Lundberg, and Su-In Lee. 2020. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *arXiv: Learning*.

Amirata Ghorbani, Abubakar Abid, and James Zou. 2018. Interpretation of neural networks is fragile.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. *ArXiv*, abs/2104.07467.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *EMNLP*.

Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *ArXiv*, abs/1711.07414.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *ACL*.

Alon Jacovi and Yoav Goldberg. 2021. Aligning Faithful Interpretations with their Social Attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *NAACL-HLT*.

Teja Kanchinadam, Keith Westpfahl, Qian You, and Glenn M. Fung. 2020. Rationale-based human-in-the-loop via supervised attention. In *DaSH@KDD*.

Klaus Krippendorff. 1980. *Content analysis an introduction to its methodology*. SAGE.

Mirko Lai, Alessandra Teresa Cignarella, D. I. H. Farías, C. Bosco, V. Patti, and P. Rosso. 2020. Multilingual stance detection in social media political debates. *Comput. Speech Lang.*, 63:101075.

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

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *SemEval@NAACL-HLT*.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Techn.*, 17:26:1–26:23.

Danish Pruthi, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Weakly- and semi-supervised evidence extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3965–3970, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2020. Stance detection benchmark: How robust is your stance detection? *ArXiv*, abs/2001.01565.

Jost Tobias Springenberg, A. Dosovitskiy, T. Brox, and Martin A. Riedmiller. 2015. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Mariona Taulé, Maria Antònia Martí, Francisco M. Rangel Pardo, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the task on stance and gender detection in tweets on catalan independence. In *IberEval@SEPLN*.

Adam Tsakalidis, Nikolaos Aletras, Alexandra I. Cristea, and Maria Liakata. 2018. Nowcasting the stance of social media users in a sudden vote: The case of the greek referendum. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*.

Jannis Vamvas and Rico Sennrich. 2020. X -stance: A multilingual multi-target dataset for stance detection. *ArXiv*, abs/2003.08385.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *LREC*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Claire Cardie. 2017. *MPQA Opinion Corpus*, page 813–832. Springer.

Zhengxuan Wu and Desmond C. Ong. 2021. On explaining your explanations of BERT: an empirical study with sequence classification. *CoRR*, abs/2101.00196.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *ACL*.

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.

B. Zhang, M. Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *ACL*.

Ye Zhang, Iain Marshall, and Byron C. Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas. Association for Computational Linguistics.

Ruiqi Zhong, S. Shao, and K. McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *ArXiv*, abs/1908.06870.

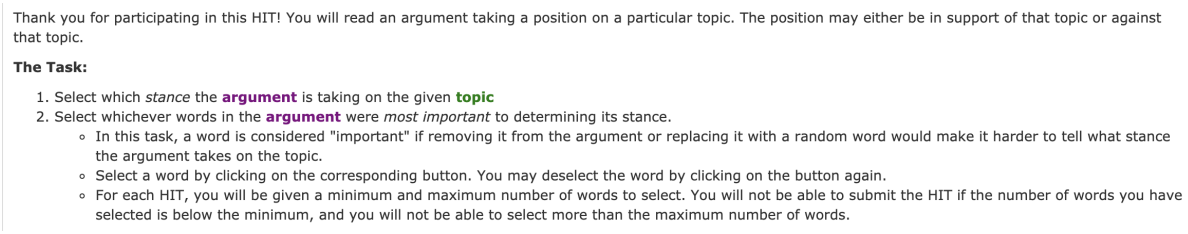


Figure 2: HIT instructions.

## A Crowdsourcing

We pay each worker $0.28 per HIT. We observe that workers spend an average of roughly 75 seconds on each HIT, excluding outliers (cases in which the worker took much longer than the other two workers for that HIT, presumably due to a workflow interruption). Only tokens containing at least one alphabetical character are selectable. For each HIT, workers are required to select at least $\text{round}(\texttt{num\_selectable}/11)$ tokens, and at most $\text{round}(\texttt{num\_selectable}/5.5)$ (where $\texttt{num\_selectable}$ is the number of selectable words in a specific HIT). See Figure 2, Figure 3, Figure 4.

**Frequently Asked Questions**

**Should I select all words in an important phrase or leave out common words like 'a', 'the', 'of', etc.?**

In most cases, selecting common words such as "a," "the," and "of" shouldn't be a *priority*. Sometimes, however, a common word is critical to a phrase. Consider the sentence, "I like to go to the community garden and take in the flowers." Removing the word 'in' from the phrase "take in" would drastically change the meaning of the sentence! Of course, if the phrase itself is not important, none of its words should be selected.

**What should I do if the paragraph takes a stance on something *related* to the topic, but not on the topic itself? For instance, if the topic is "gambling" but the argument is about gambling laws and regulations?**

Some of the topics you see will not match perfectly with their respective arguments. In such cases, make your best guess as to what the "intended topic" is and then treat that as the topic. The intended topic will always be somewhat substitutable with the given topic; for instance, if the topic is "children," then the intended topic might be "having children," but it won't be "child abuse." This is because "pro-children" could be taken to mean "pro-having-children," but it cannot be interpreted as "pro-child-abuse." Similarly, if the paragraph argues that trans fats should be banned, and the given topic is "unhealthy," then the intended topic is probably something like "trans fats are unhealthy" and the correct stance is 'For.' In the "gambling" example, it would be appropriate to assume that the intended topic is something like "unregulated gambling."

**If the topic appears in the argument, should I select it?**

That depends entirely on the context in which it appears. If you were to replace that occurrence of the topic in the paragraph with a random word or phrase, would it be significantly harder to figure out the author's stance? If so, then you should select the topic. If not, you should leave it unselected.

**If the author presents a statement for the purpose of rejecting it, should I select the important words in that statement?**

Only if the rest of the paragraph isn't enough to determine the author's stance on the topic. If you do select words in a part of the paragraph that goes against how the author feels, make sure you also select words that indicate the author's disagreement with that piece of text.

Figure 3: HIT FAQ.

**Topic:** vaccination

**Argument:** We still haven't addressed the problem of children who come from other countries into America and have not had the advantage of early childhood vaccinations. We are seeing cases of measles, mumps, etc. increase, and this cannot be all the fault of those who have lived with and been in contact daily in this country since birth with their peers who have been vaccinated. I'm old, and when in 1st grade, I contacted German measles. NOT from someone in my class and, in fact, no one seemed to know where I contacted it. I then proceeded to infect almost all my entire class apparently. My point in telling this story is that we cannot know exactly where a child gets his contact, and therefore IMO opinion, vaccinations should be mandatory for every child entering schools from pre-K up. Until everyone - from the child who comes with their parents to America to the kiddo next door, we have to be consistent.

**Question 1.** What stance does the argument take toward the topic?

○ **For**

○ **Against**

**Question 2.** Which words in the argument are most important to identifying its stance on "vaccination?" Please select 17-34 words.

We | still | have | n't | addressed | the | problem | of | children | who | come | from | other | countries | into | America | and | have | not | had | the | advantage | of | early | childhood | vaccinations | . | We | are | seeing | cases | of | measles | , | mumps | , | etc | . | increase | , | and | this | can | not | be | all | the | fault | of | those | who | have | lived | with | and | been | in | contact | daily | in | this | country | since | birth | with | their | peers | who | have | been | vaccinated | . | I | 'm | old | , | and | when | in | 1st | grade | , | I | contacted | German | measles | . | NOT | from | someone | in | my | class | and | , | in | fact | , | no | one | seemed | to | know | where | I | contacted | it | . | I | then | proceeded | to | infect | almost | all | my | entire | class | apparently | . | My | point | in | telling | this | story | is | that | we | can | not | know | exactly | where | a | child | gets | his | contact | , | and | therefore | IMO | opinion | , | vaccinations | should | be | mandatory | for | every | child | entering | schools | from | pre | - | K | up | . | Until | everyone | - | from | the | child | who | comes | with | their | parents | to | America | to | the | kiddo | next | door | , | we | have | to | be | consistent | .

2 words selected

Figure 4: Example HIT.

## B Other Diagnostic Properties

When assessing GI and MAW as explainability techniques for our models, we choose not to consider four of the five diagnostic properties proposed by Atanasova et al. (2020). *Agreement with human annotations (HA)* is not necessarily a desirable property, as it is little more than an indication of how *convincing* an attribution method is to humans. Note that we compute HA in the form of rationale loss (§5.4), but do so as a way of evaluating attributions themselves, as opposed to attribution methods. Confidence Indication (CI) does not apply to MAW, as attention weights do not differ by class for a fixed input. The authors' metric for Rationale Consistency (RC) requires the assumption that models with similar reasoning paths have similar activation maps, an assumption we believe is flawed primarily on account of architectural symmetry. Lastly, we believe that the proposed metric for Dataset Consistency (DC) would not be meaningful for our dataset, as the degree of similarity between different arguments in VAST is extremely low.

## C Visualizing Attributions

To visualize model attributions for figures, human evaluation, and rationale error analysis, we map the attribution score for each token to a new score of 0 (unselected), 0.5 (selected but only moderately important), or 1 (selected and very important). We perform this mapping using the following procedure, which takes parameters $k$ and $\epsilon$:

1. Rank the attribution scores for the input sequence in descending order, and let `k_score` be the score of the $k$-th item.
2. Assign all tokens with score $>$ `k_score` $+ \epsilon$ a new score of 1.
3. Assign all tokens with score $<$ `k_score` $- \epsilon$ a new score of 0.
4. Assign all other tokens a new score of 0.5.

For an argument of length $m$, we set $k = m/8$. We let $\epsilon = .05 * \texttt{max\_att}$, where `max_att` is the maximum of the original attribution scores for the argument. We obtain these values through trial-and-error on training examples, with the subjective goal of achieving visuals that contain a meaningful number of "moderately important" and "very important" words while reflecting stratifications in the original attribution scores. We take the new score of a multi-token word to be the maximum new score over its subword tokens.

| $\lambda$ | full | reduced$_{25}$ | reduced$_{10}$ |
|---|---|---|---|
| 0 | .726 | .710 | .699 |
| 16,384 | *n/a* | .701 | **.703** |
| 32,768 | .712 | .698 | .702 |
| 49,152 | **.726** | **.712** | .695 |
| 65,536 | .711 | .703 | *n/a* |

Table 10: Dev set results for various $\lambda$ (evenly spaced by $2^{14} = 16384$) in each of the data settings. $\lambda = 0$ indicates that our attribution prior was not applied. A single random seed was used. *n/a* indicates that the trial was not performed.

## D Choosing $\lambda$

See Table 10.

## E Variance Across Trials

See Table 11, Table 12.

| | | All | Zero | Few |
|---|---|---|---|---|
| full | `base` | 6.5 | 2.2 | 29.1 |
| | `p-b:g` | 28.0 | 23.4 | 31.0 |
| reduced$_{25}$ | `BT-j` | 15.9 | 24.6 | 8.0 |
| | `base` | 8.2 | 6.6 | 11.2 |
| | `p-b:g` | 10.2 | 3.2 | 20.1 |
| reduced$_{10}$ | `BT-j` | 11.4 | 16.9 | 7.4 |
| | `base` | 103.0 | 120.1 | 90.1 |
| | `p-b:g` | 6.9 | 5.4 | 4.9 |

Table 11: Variance across trials for Combined F1 results ("All") reported in Table 3, multiplied by $10^5$. `p-b:g` refers to `prior-bin:gold` (`BT-j`, `BERT-joint`).

## F Misc.

Our model consists of $109,917,780$ parameters. Training on the `full` train set using our proposed attribution prior takes 11 hours and 16 minutes using two Tesla T4 GPUs.

| | | *Imp* | | *mlT* | | *mlS* | | *Qte* | | *Sarc* | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MAW | GI | MAW | GI | MAW | GI | MAW | GI | MAW | GI |
| `base` | I | 5.2 | 2.5 | 10.2 | 5.6 | 10.2 | 6.0 | 7.1 | 5.0 | 90.2 | 30.7 |
| | O | 6.8 | 19.2 | 1.0 | 0.0 | 11.3 | 4.6 | 100.9 | 90.1 | 5.6 | 0.3 |
| `prior-` | I | 0.0 | 0.0 | 0.1 | 1.5 | 0.0 | 5.6 | 0.0 | 4.0 | 0.5 | 3.2 |
| `bin:gold` | O | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 | 16.9 | 0.0 | 40.5 | 0.0 | 3.6 |

Table 12: Variance across trials for $\sqrt{\lambda \mathcal{L}_p}$ reported in Table 9, multiplied by $10^5$.