Prompting for explanations improves Adversarial NLI. Is this true? {Yes} it is {true} because {it weakens superficial cues}

Pride Kavumba[1,2] **Ana Brassard**[2,1] **Benjamin Heinzerling**[2,1] **Kentaro Inui**[1,2]

[1]Tohoku University [2]RIKEN AIP
kavumba.pride.q2@dc.tohoku.ac.jp
{ana.brassard, benjamin.heinzerling}@riken.jp
kentaro.inui@tohoku.ac.jp

Abstract

Explanation prompts ask language models to not only assign a label to a given input, such as *entailment* or *contradiction* in natural language inference (NLI) tasks, but also to generate a free-text explanation that supports this label. While explanation prompts originally introduced aiming to improve model interpretability, here we show that they also improve robustness to superficial cues. Compared to prompting for labels only, explanation prompting shows stronger performance on adversarial NLI benchmarks, outperforming the state of the art on ANLI, Counterfactually-Augmented NLI, and SNLI-Hard datasets. Analysis suggests that the increase in robustness is due to a reduction in the association strength between single tokens and labels, i.e., explanation prompting weakens superficial cues. More specifically, we find that single tokens that are highly predictive of the correct answer in the label-only setting become uninformative when the model also has to generate explanations.

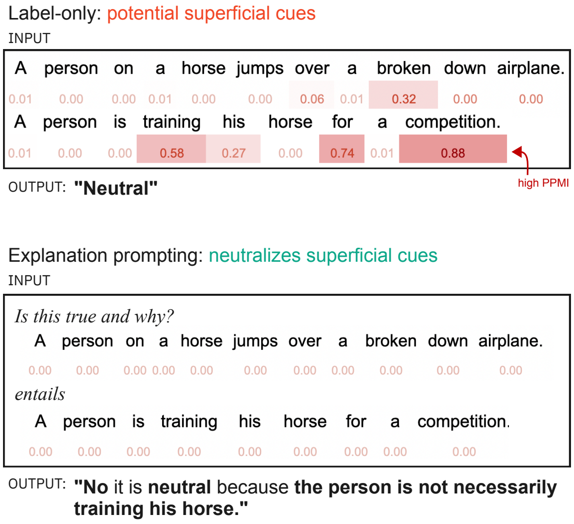

Label-only: potential superficial cues
INPUT
A person on a horse jumps over a broken down airplane.
0.01 0.00 0.00 0.01 0.00 0.00 0.06 0.01 0.32 0.00 0.00
A person is training his horse for a competition.
0.01 0.00 0.00 0.58 0.27 0.00 0.74 0.01 0.88
high PPMI
OUTPUT: "Neutral"
Explanation prompting: neutralizes superficial cues
INPUT
Is this true and why?
A person on a horse jumps over a broken down airplane.
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
entails
A person is training his horse for a competition.
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
OUTPUT: "No it is neutral because the person is not necessarily training his horse."

Figure 1: When models only have to predict class labels (top), some words in the input can become superficial cues, as indicated by high pointwise mutual information (shown in red) between words and class labels. With *explanation prompting* (bottom) the added requirement of generating explanations renders such shortcuts ineffective.

1 Introduction

Explanation prompting requires language models to not only assign a particular label to a given input (henceforth: label-only prompting), but also to generate an explanation that supports this label. For example, given the natural language inference (NLI; Bowman et al., 2015) premise "A soccer game with multiple males playing" and the hypothesis "Some men are playing a sport", in label-only prompting the model only has to generate a label such as *entailment*. With explanation prompting, the model has to generate not only the label but also an *explanation* that supports this label, such as "It is true because playing soccer is playing a sport".

While explanation prompting was originally proposed for improving model interpretability (Narang et al., 2020), here we explore a different advantage: improved model performance on adversarial benchmarks. Created in response to the discovery of superficial cues in many common datasets, adversarial benchmarks are designed to give a more realistic estimate of model performance. Non-adversarial benchmarks such as SNLI (Bowman et al., 2015) can contain superficial cues, i.e., single tokens that are predictive of the correct label and hence allow models to achieve high scores by taking "shortcuts" instead of acquiring and employing the capabilities intended by the task designers (Gururangan et al., 2018a; McCoy et al., 2019; Poliak et al., 2018; Niven and Kao, 2019; Sugawara et al., 2018; Schuster et al., 2019a; Kavumba et al., 2019). In contrast, adversarial benchmarks are created in a way that reduces or completely eliminates superficial cues, thus forcing models to solve tasks in the intended and generally more difficult manner.

Findings of the Association for Computational Linguistics: EACL 2023, pages 2165–2180
May 2-6, 2023 ©2023 Association for Computational Linguistics

In this work, we investigate the benefits of explanation prompting through the lens of adversarial benchmarks. Concretely, we finetune pretrained language models on natural language inference datasets with explanation prompting and compare performance to label-only prompting. We find that explanation prompting improves performance across four adversarial NLI datasets and two non-adversarial NLI datasets (§5). Improvements are consistent across model architectures, model sizes, and prompt variations. Further analysis reveals that both the specific verbalization of the label ("*Yes, it is* `{label}` because...") and the relation between explanation and label are important for model performance (§6). Finally, we verify that explanation prompting models do not rely on the kind of superficial cue that allows taking shortcuts in the label-only setting (§6). Source code is available at `github.com/pkavumba/explanation-prompting`.

2 Background and Related Work

2.1 Superficial Cues

In the original natural language inference setting, as exemplified by SNLI (Bowman et al., 2015), models are trained to assign a label, such as *entailment* or *contradiction*, to a given input. While models quickly achieved high evaluation scores, a line of research starting with Gururangan et al. (2018a) found that SNLI and other datasets contain superficial cues that models can exploit instead of learning the task as intended (Poliak et al., 2018; McCoy et al., 2019; Niven and Kao, 2019; Schuster et al., 2019a; Kavumba et al., 2019; Wang and Culotta, 2021; Srivastava et al., 2020; Wang et al., 2019b,c). For example, in SNLI, negations such as "not" are strongly associated with the *contradiction* label (Gururangan et al., 2018a). A model that predicts *contradiction* when the input contains the token "not" will achieve a high evaluation score without acquiring any capability to perform actual natural language inference. Having learned to rely on such shortcuts (Geirhos et al., 2020), models will be "right for the wrong reasons" (McCoy et al., 2019) on data that contains superficial cues, but will perform worse on data that does not.

There are several approaches to mitigate superficial cues. A direct countermeasure is to remove them from existing datasets and to take care not to introduce superficial cues when creating new datasets. The two dominant methods to do so are removal of easy samples via adversarial filtering (Zellers et al., 2018, 2019; Sakaguchi et al., 2020; Bras et al., 2020; Nie et al., 2020) and augmentation with counterfactual examples that neutralize the association between existing superficial cues and labels (Kavumba et al., 2019; Schuster et al., 2019b; Kaushik et al., 2020). A complementary line of work aims to prevent models from relying on superficial cues, for example via adversarial training (Belinkov et al., 2019; Stacey et al., 2020, 2021) and adversarial attacks (Wang et al., 2019a; Liu et al., 2020; Zhu et al., 2020; Wang et al., 2021). Adversarial approaches suffer from drawbacks such as a more complex training scheme and higher computational costs. Another approach is multi-task training. Camburu et al. (2018) propose a "predict-and-explain" multi-task setup in which one model first predicts a label and a second model generates a free-form explanation for this label However, this setup turns out to slightly degrade performance.

In this work, we study *explanation prompting* as a method for reducing the impact of superficial cues. While this form of prompting was originally introduced to enhance model interpretability (Narang et al., 2020), our work is most closely related to Chen et al. (2022), who studied the robustness of rationale models (Lei et al., 2016; Bastings et al., 2019; DeYoung et al., 2020) to adversarial attacks. Rationale models operate in a two-step "rationalize-then-predict" manner, where the model first selects a pertinent subset of the input, called a *rationale*, and then predicts a label given this rationale. Stacey et al. (2021) investigates using human annotated rationales for supervising attention mechanism. Their goal is to increase the attention given to annotated rationales.

2.2 Explanation Prompting

Explanation prompting requires models not only to predict a class label but also to provide an explanation of why that label is the correct answer. Previous work has explored explanation prompting as a way to improve model interpretability. Wiegreffe et al. (2021) analyzed the faithfulness of explanations obtained via explanation prompting. Since high-quality explanation are expensive to create and not available in large quantities, Marasovic et al. (2022) compare methods for generating high-quality explanations in limited data regimes, whereas Wiegreffe et al. (2022) investigate the feasibility of using large language models such as GPT-3 (Brown et al., 2020) to automatically generate large amounts of explanations. In contrast to this strand of research, we use free-text explanations not to improve model interpretability, but to improve model robustness in adversarial settings.

3 Explanation Prompting for Adversarial NLI

In the original natural language inference setting, one trains a classifier to label the relationship between a *premise* and a *hypothesis* as *entailment*, *neutral*, or *contradiction*. When using a generative language model to generate a label and an explanation supporting the label, the task turns from classification into what we refer to as *explanation prompting*. Turning the original NLI instances into input-output pairs suitable for a generative language model necessitates choosing a *verbalizer*[1] that converts premise, hypothesis, label, and answer into an input prompt an target output, e.g.:

- *Input:* Is this true and why? `{premise}` implies `{hypothesis}`
- *Output:* `{Yes or No}` it is `{label}` because `{explanation}`

Note that the label prediction process is further broken down into two steps: first, the model must provide a *binary* answer to the question (in this case, whether the statements are entailed), then give the exact label. That is, the output starts with *"**Yes** it is ..."* for `Entailment`, or *"**No** it is ..."* for `Neutral` and `Contradiction`. We refer to this as *multi-step verbalizing*. This is a deviation from previous work that utilized single-step or single-word verbalizers. For example, `entailment` is often verbalized as *yes*, while `contradiction` and `neutral` are verbalized as *no* and *maybe*, respectively (Schick and Schütze, 2021a,b). Finally, the output is completed with a free-text explanation supporting the label.

4 Experimental Setup

4.1 Datasets

We compare label-only and explanation prompting on six NLI datasets.

e-SNLI (Camburu et al., 2018) extends SNLI (Bowman et al., 2015) with crowdsourced free-form explanations and annotated salient spans.

Adversarial NLI (ANLI) (Nie et al., 2020) was created in an iterative, adversarial process where, in each iteration, human annotators create examples that a given model does not label correctly, which are then used to train a stronger model.

SNLI Hard (Gururangan et al., 2018a) is a filtered version of the SNLI test set and contains only instances that could not be labeled correctly by a model given only the hypothesis as input.

NLI Diagnostic (Wang et al., 2018) was carefully constructed to evaluate capabilities related to commonsense knowledge, logical reasoning, predicate-argument structures, and lexical semantics.

Heuristic Analysis for NLI Systems (HANS) (McCoy et al., 2019) was created to analyze and prevent several kinds of shortcuts found in prior NLI datasets, such as lexical overlap between premise and hypothesis.

Counterfactually-Augmented NLI (Counter-NLI) (Kaushik et al., 2020) augments a subset of SNLI with counterfactual instances, which were obtained by editing either the premise or hypothesis so that a counterfactual, i.e., different than the original, label becomes true. Models relying on superficial cues will perform well on original SNLI instances, but poorly on counterfactual ones.

4.2 Models and Training Details

The two main models selected for our comparison are T0 (Sanh et al., 2021) and T5-3B (Raffel et al., 2020). We chose these two models based on their good reported performance on NLI datasets while involving comparably low computational costs, which we further reduced by finetuning all models only on a third of the available e-SNLI and ANLI training data. Thus, test results on e-SNLI, ANLI, and SNLI Hard can be considered in-domain tests and results on the remaining datasets out-of-domain tests. Further training details and hyperparameter settings are given in Appendix B.

5 Results

Does explanation prompting improve robustness to adversarial attacks? *Yes.*

For both T0 and T5-3B, training with explanation prompting improved performance over label-only prompting on nearly all datasets, surpassing the reported state of the art on e-SNLI, SNLI-Hard,

[1] Verbalizer details are given in Appendix A.

Dataset	Subset	Current SOTA	T5-3B		T0 (11B)	
			Label-only	Explanation prompting	Label-only	Explanation prompting
e-SNLI	-	92.3	91.7	**95.1**	91.0	**91.9**
SNLI Hard	Hard	80.2	84.0	**89.7**	83.0	**84.5**
ANLI	R1	75.5	74.9	**81.8**	69.6	**75.6**
	R2	51.4	58.9	**72.5**	53.7	**60.6**
	R3	49.8	57.9	**74.8**	55.0	**59.9**
HANS	Lex	94.1	**94.2**	**94.2**	**97.9**	95.9
	Sub	46.3	**46.3**	30.3	20.5	**37.9**
	Cons	38.5	**38.6**	17.1	24.3	**53.9**
Counter-NLI	RP	54.3	69.6	**83.0**	66.5	**69.2**
	RH	74.3	88.9	**93.5**	**87.9**	87.4
	RP&RH	64.3	79.3	**88.3**	77.2	**78.3**
NLI Diagnostic	Know	53.9	58.8	**76.4**	58.8	**59.9**
	Logic	58.7	63.7	**73.9**	60.7	**64.5**
	LS	66.5	69.6	**79.3**	63.0	**70.4**
	PAS	69.9	73.1	**80.9**	70.8	**72.4**

Table 1: Results by T5-3B and T0 (11B) models trained with label-only prompting and Explanation prompting. Current state-of-the-art results on each dataset are reported from: WT5 (Narang et al., 2020), BERT-Sup-ATT (Stacey et al., 2021), InfoBERT (Wang et al., 2021), RoBERTa-AFLITE (Bras et al., 2020), BERT (Kaushik et al., 2020), and RoBERTa-AFLITE (Bras et al., 2020), respectively. Note that T5-3B and T0 are trained with different batch sizes and sequence lengths, so the results are not comparable (§ 5).

ANLI, and Counterfactually-Augmented NLI (Table 1). For example, on the three ANLI subsets T5-3B achieves accuracies of 81.8%, 72.5%, and 74.8% with explanation prompting, compared to much lower accuracies of 74.9%, 58.9% and 57.9% with label-only prompting. Furthermore, since e-SNLI does not contain any adversarially chosen "hard" instances, strong results on this dataset show that explanation prompting does not necessarily hurt performance on datasets with superficial cues. Overall T5-3B achieves higher performance despite its smaller size, but this is due to T0 using a quarter of the batch size and sequence length of that used for T5 due to memory limitations.

The HANS dataset remains the most challenging dataset, indicating that the models may still be susceptible to such adversarial attacks. Surprisingly, the use of explanation prompting actually leads to *degraded* performance for certain subsets, such as the lexical overlap for T0, and subsequences and constituents for T5-3B. This discrepancy warrants further investigation, which we leave for future work.

On all other datasets, explanation prompting models show clear improvements over label-only models in both in-domain and adversarial out-of-domain settings. This demonstrates that using explanation prompting generally enhances the model's robustness to adversarial attacks and improves the overall NLI prediction performance.

	Full	H-only	Δ
Label-only prompting	87.2	63.7	-23.5
Explanation prompting	90.9	33.1	-57.8
Random baseline	33.3	33.3	-

Table 2: The average prediction accuracy of T0 models on e-SNLI when trained with the full input compared to the hypothesis-only setting (H-only), which allows the models to solely rely on superficial cues to make accurate predictions. The explanation prompting-trained model's performance degraded to random performance, implying that it did not learn to make use of superficial cues.

6 Discussion

In this section we vary experimental settings and conduct ablations in order to provide a more detailed analysis of how explanation prompting impacts NLI performance. Unless stated otherwise, reported results are obtained by finetuning T0 on 20K randomly-sampled instances from e-SNLI and averaging prediction accuracies from three runs

Explanation	Accuracy	BLEU
None	88.4	-
Random characters	21.00	0.02
Random words	0.00	0.90
Low-sim. sentences	0.04	0.03
High-sim. sentences	59.1	1.73
Original (e-SNLI)	**91.6**	36.1

Table 3: Impact of explanation content. Accuracies are shown for the e-SNLI dev set, averaged over three T0 models finetuned with different random seeds. Random characters and words or unrelated explanations significantly reduce performance, indicating that the models did not rely on superficial cues. Original explanations outperform extracted sentences with high similarity, demonstrating the benefit of related explanations. We also report BLEU scores with respect to the original explanations.

with different random seeds. Training details are given in Appendix B.

Does explanation prompting prevent models from exploiting superficial cues? *Yes.*

To see if models still exploit superficial cues, we employ the hypothesis-only setting of Gururangan et al. (2018b). Since the missing premise makes the task impossible, any performance above random chance can be ascribed to models picking up on superficial cues. After training one model with label-only prompting and one with explanation prompting(Table 2), we observe that the label-only model considerably exceeds random chance (63.7% compared to 33.3%). In contrast, the explanation prompting model does not exceed random chance, indicating that explanation prompting is not conducive to shortcut learning.

Do the explanations need to be related to the input? *Yes.*

To check if the content of the target explanation matters we replace it with unrelated text ranging from completely random characters to similar but unrelated sentences and find that explanation prompting with the original explanations still performs best (Table 3). Specifically, we choose the following target "explanations": (i) random characters, (ii) random words, (iii) sentences extracted from the BookCorpus (Zhu et al., 2015) with *low* similarity with the input, and (iv) sentences extracted from the BookCorpus with *high* similarity.[2] All similarities are computed with Sentence-BERT (Reimers and Gurevych, 2019). We also compare to label-only prompting (*None* row in Table 3). Table 3 shows the mean prediction accuracy scores on the development set of e-SNLI over three random seeds. Performance degrades with random explanations or sentences extracted from BookCorpus, confirming that training the model to predict explanations improves adversarial robustness.

Does *Multi-step* verbalizing have an effect on the model performance? *Yes.*

A binary decision step may seem like a small change in prompt format, however, we found that this added step has partial merit to the improvement in performance. To verify this, we train label-only and explanation prompting models using *single-step*-verbalized prompts (*{label} because...*") and ones using *multi-step*-verbalized prompts (*"Yes it is {label} because..."*), both with an explanation (+Explain column in Table 5) and without added explanations (Label-only column in Table 5). The results show the prediction accuracy averaged over three random seeds. In both label-only and explained settings, adding a multi-step verbalizer brings an improvement over the single-step version.

Are the models sensitive to prompt wording in the input? *No.*

Previous work has demonstrated that language models can be very sensitive to the prompts (Schick and Schütze, 2021a,b; Brown et al., 2020). To examine this, we conduct experiments on five diverse crowdsourced prompts obtained from the Prompt Source project (Bach et al., 2022). For each model, we run three separate experiments using three different random seeds. We report the average accuracy across all five prompts on the development set of e-SNLI and Counterfactually-Augmented NLI. Due to resource constraints, we use T5-3B, a smaller model than T0, for these experiments. Furthermore, we limit the number of instances used from e-SNLI to twenty thousand randomly selected examples.

The results presented in Table 4 demonstrate that models trained with explanation prompting outperform those trained with label-only prompting across all five prompts in terms of accuracy (see Table 8 in Appendix C for results with different models). For instance, the explanation prompt-

[2] We use the BookCorpus instead of sampling random explanations to avoid accidentally sampling valid explanations.

Dataset	Prompt ID 1	2	3	4	5	Mean$_{(stddev)}$
e-SNLI	91.5 / **94.4**	91.6 / **94.7**	91.8 / **94.6**	91.6 / **94.5**	91.8 / **94.4**	91.7$_{(0.1)}$ / **94.5**$_{(0.1)}$
CNLI (RP)	70.8 / **82.5**	73.0 / **83.8**	71.5 / **83.0**	72.1 / **83.0**	70.1 / **82.8**	71.5$_{(1.1)}$ / **83.0**$_{(0.5)}$
CNLI (RH)	82.0 / **92.3**	82.8 / **93.0**	81.9 / **92.8**	82.0 / **92.2**	83.1 / **92.5**	82.4$_{(0.6)}$ / **92.6**$_{(0.3)}$
CNLI (RP&RH)	76.4 / **87.4**	77.9 / **88.4**	76.7 / **87.9**	77.0 / **87.6**	76.6 / **87.7**	76.9$_{(0.6)}$ / **87.8**$_{(0.4)}$

Table 4: Prompt sensitivity on the development set of e-SNLI and Counterfactually-Augmented NLI (CNLI). Values are accuracy of label-only/explanation prompting-trained T5-3B models averaged over three random seeds. Besides the consistently higher performance of the explanation prompting setting, the lower standard deviation indicates greater stability w.r.t. prompt format.

	Label-only	+Explain (e-SNLI)
Single-step	87.2	90.9
Multi-step	88.4	**91.6**

Table 5: Comparing the effects of single-step ("`[label]` *because ...*") and multi-step ("*Yes/no, it is* `[label]` *because ...*") verbalizing on T0 prediction accuracy, both with an explanation (+Explain) and without an explanation (Label-only) in the model output. Multi-step verbalizing improved the accuracy in both cases, with and without explanation, and the added task of providing an explanation (+Explain) further enhanced performance.

ing model achieves an overall average accuracy of 94.5% on e-SNLI compared to 91.7% for the label-only model. Additionally, the explanation prompting model exhibits better accuracy on all the individual prompts on the adversarial Counterfactually-Augmented NLI with an overall average accuracy of 83.0% on the revised premise (RP), 92.6% on the revised hypothesis (RH), and 87.8 on RP&RH, compared to 71.5%, 82.4%, and 76.9% respectively for the label-only model. The lower standard deviations for explanation prompting also indicate higher stability across all prompts.

Are the results dependent on the architecture/size of the model employed? *Yes.* To study the impact of model size on performance, we repeat the experiments using six models ranging from 60 million to 11 billion parameters. These models comprise two versions of BART (Lewis et al., 2020) with 125M and 400M parameters, as well as three variants of T5 (Raffel et al., 2020) with 60M, 770M, and 3B parameters, and T0 with 11B parameters. The results, as depicted in Figure 3 for ANLI, indicate a clear correlation between model size and performance, with larger models demonstrating improved results. It is worth noting that the highest achieved performance by T5-3B, exceeds that of the larger T0 model. This is due to the fact that T0 is trained using only a quarter of the batch size and the sequence length used for all other models, resulting in reduced performance. Comprehensive results for all models and datasets are presented in Table 6. For more information on the training details, see Appendix B.

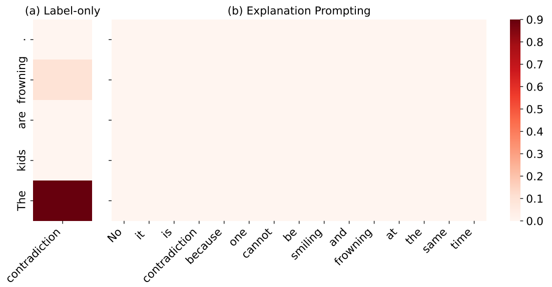


Figure 2: Positive Pointwise Mutual Information (PPMI) statistics for hypothesis words and labels with and without explanation prompting. Words in the hypothesis are strongly associated with the predict-only labels. With explanation prompting, the association between the hypothesis words and the labels drops to a zero. For example, while the negative word *frowning* is strongly associated with *contradiction* label, the association is eliminated with explanation prompts. The figure only shows the hypothesis because superficial cues are from the hypothesis, not the premise.

Does explanation prompting weaken the association between word-level superficial cues and labels? *Yes.* To investigate the impact of explanation prompting on the association between input words and their corresponding output labels in the training set, we compare the positive pointwise mutual information (PPMI) between them in both label-only and explanation prompting settings:

		Current	T5-Small (60M)		BART-Base (125M)		BART-Large (400M)		T5-Large (770M)		T5-3B (3B)		T0* (11B)	
		SOTA	LP	EP	LP	EP	LP	EP	LP	EP	LP	EP	LP	EP
e-SNLI		92.3	82.4	88.8	88.7	92.1	90.4	93.8	90.9	94.4	91.7	**95.1**	91.0	91.9
SNLI Hard		80.2	68.5	82.2	78.1	84.3	81.5	84.9	82.1	88.7	84.0	**89.7**	83.0	84.5
ANLI	R1	75.5	46.5	52.5	56.8	53.0	64.9	65.9	66.1	77.2	74.9	**81.8**	69.6	75.6
	R2	51.4	37.6	56.4	41.5	50.3	44.4	57.1	49.2	67.8	58.9	**72.5**	53.7	60.6
	R3	49.8	40.4	59.1	40.9	54.0	46.5	59.6	49.4	68.0	57.9	**74.8**	55.0	59.9
HANS	Lex	94.1	2.6	0.0	71.2	69.6	85.0	90.2	82.9	81.3	94.2	94.2	**97.9**	95.9
	Sub	46.3	2.2	0.0	43.2	54.1	27.3	**63.7**	35.6	27.6	46.3	30.3	20.5	37.9
	Cons	38.5	2.5	0.0	34.7	51.9	22.4	**63.8**	19.6	9.9	38.6	17.1	24.3	53.9
Counter-NLI	RP	54.3	54.1	75.6	59.8	74.9	66.1	77.3	67.8	82.3	69.6	**83.0**	66.5	69.2
	RH	74.3	78.4	86.5	82.9	87.8	85.3	87.4	86.5	92.4	88.9	**93.5**	87.9	87.4
	RP&RH	64.3	66.3	81.1	71.3	81.3	75.7	82.3	77.1	87.3	79.3	**88.3**	77.2	78.3
NLI Diagnostic	Know	53.9	34.5	58.8	41.2	60.2	57.4	70.4	54.9	65.8	58.8	**76.4**	58.8	59.9
	Logic	58.7	45.3	59.6	45.6	67.0	54.9	67.0	57.4	70.3	63.7	**73.9**	60.7	64.5
	LS	66.5	49.5	63.3	49.2	62.2	62.2	69.6	63.9	76.1	69.6	**79.3**	63.0	70.4
	PAS	69.9	58.0	69.3	55.7	65.3	67.9	66.7	71.0	76.4	73.1	**80.9**	70.8	72.4

Table 6: Average prediction accuracy over three random seeds by models of increasing size trained with label-only (LP) and explanation prompting (EP). Current state-of-the-art results on each dataset are reported from: WT5 (Narang et al., 2020), BERT-Sup-ATT (Stacey et al., 2021), InfoBERT (Wang et al., 2021), RoBERTa-AFLITE (Bras et al., 2020), BERT (Kaushik et al., 2020), and RoBERTa-AFLITE (Bras et al., 2020), respectively. *Note that the T0 models were trained using only a quarter of the batch size and half the sequence length used for all other models due to computational limitations. This may be the cause for weaker performance compared to the smaller T5-3B models.

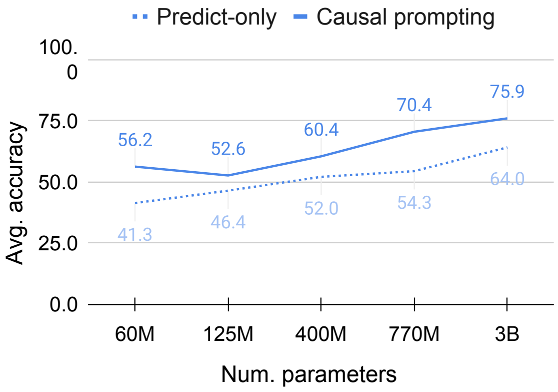


Figure 3: Average accuracy on ANLI depending on model size with label-only and explanation prompt training, respectively. See Appendix B for a comprehensive overview of performance, including on other datasets.

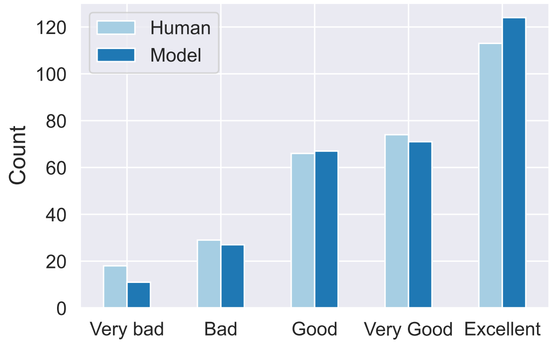


Figure 4: Crowd-sourced comparison of human-written explanations and those generated by a model. Overall, the crowd workers found that the model-generated explanations were either comparable to or better than the human-written ones.

$$PPMI(w,l) = max(log\frac{p(w,l)}{p(w,)p(,l)}, 0)$$

Where w represents the input word and l represents the output label. The PPMI analysis enables us to determine which words have a strong association with specific output labels, and how the use of explanation prompts modifies these associations. Following Gururangan et al. (2018b), we use add-100 smoothing in our PPMI calculations to highlight the input words that exhibit the strongest associations with the output labels.

The results, as presented in Figure 2, show that explanation prompts weaken the association between the input words and the output labels. For instance, in the absence of explanation prompting, the negative word *frowning* had a strong association with the label *contradiction*. However, when explanation prompting is used, this association is diminished from 0.1 to around 0. It's worth mentioning that only the hypothesis words are shown in Figure 2 as the superficial cues are mainly present in the hypothesis, not the premise. These findings align with the results obtained from the hypothesis-only model, as presented in Table 2.

Are the explanations generated by explanation prompting models plausible? *Yes.*

While interpretability is not the primary objective of this study, we conduct a human evaluation of the model's generated explanations to assess its performance within the intended framework. We use Amazon Mechanical Turk to gather assessments from 100 randomly selected T0 instances. To make the task easier for the crowd workers, we simplify it to only include two labels: `Entailment` and `non-entailment`. Instances labeled as `Neutral` or `Contradiction` are considered as `non-entailment`. We present the crowd workers with the gold label and request an evaluation of the quality of the *explanation* using a five-point Likert scale, ranging from "very bad" to "extremely good". We gather three ratings per instance. Each Human Intelligence Task (HIT) features three explanations: the gold explanation authored by a human, the model-generated explanation, and an "attention check" explanation with a known expected annotation. The "attention check" explanation is included to ensure the quality of the annotations provided. This "attention check" explanation is randomly selected and unrelated to the *premise* and *hypothesis*, and therefore is expected to receive a lower score compared to the human-authored explanations that have already been validated by other crowd workers in previous work (Camburu et al., 2018). If the "attention check" explanation receives a high score or an equal score to a human-authored explanation, the HIT is flagged for review. Additionally, to prevent the use of simple heuristics, such as assuming that the last explanation is always the low-scoring one, the order in which the explanations are presented to the annotator is randomly shuffled during each HIT. An example of an NLI instance with an "attention check" explanation is shown below:[3]

- **Premise:** A man stands by an animal rights sign at an outdoor event.
- **Hypothesis:** A man is standing inside of his house
- **Human:** an outdoor event is not in his house
- **Generated:** The man cannot be standing inside of his house and at an outdoor event at the same time.
- **Attention Check:** It cannot be inferred that the young woman is an artist or that she is be finished soon.

The results of this evaluation are shown in figure 4. On the whole, crowd workers found model-generated explanations to be comparable to or better than human-written ones.[4]

This result suggests that the crowd workers found the generated reasons to be of similar quality to the human-authored reasons. This indicates that the model learns important features of the input data and is able to use them effectively to generate reasonable explanations.

7 Conclusions

In this study, we examined the influence of causal prompting on the adversarial robustness of natural language processing models. Our results indicate that using causal prompts can improve a model's robustness to adversarial attacks. We also explored the performance of our models under various modified and ablated settings and found that explanation prompting-trained models (i) no longer rely on superficial cues, (ii) benefit most from both causally related explanations and multi-step verbalization, and (iii) are robust to differences in the input prompts. In addition, we observed that performance increases with model size and that the use of the explanation prompting format reduces the association between input words and output labels. Finally, human evaluation showed that the models generated plausible explanations.

Limitations

Explanation prompting requires datasets annotated with explanations, which may not always be available and it can be costly to collect explanations in a quantity suitable for model training or finetuning. Therefore, applying this method to datasets without explanations may be difficult.

Additionally, our analysis and evaluation are limited to English language benchmarks. Although we anticipate the method to be transferable to other languages, this requires further investigation.

Finally, experiments showed that training with explanation prompting did not improve the performance of T5 variants on the "Subsequence" and "Constituent" subsets of the HANS dataset. It is

[3] The crowdsourcing study form can be found in the appendix D.

[4] Refer to Appendix E for some example explanations.

currently unclear why all variants of BART performed better than random baselines, while T5 variants did not. Possible explanations include differences in training data, model architecture, and optimization goal, but this discrepancy requires further investigation.

Ethics Statement

In our human evaluation of the explanation quality, we took steps to ensure fair treatment of the crowd workers. To do this, we internally conducted experiments to determine the average completion time for one Human Intelligence Task (HIT). Our goal was to pay a fair wage of at least $20 per hour to all participants involved in the study. As a result, all crowd workers received fair compensation for their work, and no HITs were rejected.

Acknowledgements

We thank the reviewers for their constructive feedback. We also thank all the crowd workers who have contributed to this research by rating the explanations. This work was supported by JST CREST Grant JPMJCR20D2 and JSPS KAKENHI Grant JP21K17814, Japan.

References

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 256–262, Minneapolis, Minnesota. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805, Seattle, United States. Association for Computational Linguistics.

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

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018a. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018b. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually augmented data. *International Conference on Learning Representations (ICLR)*.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When choosing plausible alternatives, clever hans can be clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models.

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.

Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. 2021. Zero-offload: Democratizing billion-scale model training.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019a. Towards debiasing fact verification models. 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 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019b. Towards debiasing fact verification models. 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 3410–3416, Hong Kong, China. Association for Computational Linguistics.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9109–9119. PMLR.

Joe Stacey, Yonatan Belinkov, and Marek Rei. 2021. Supervising model attention with human explanations for robust natural language inference.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. There is strength in numbers: Avoiding the hypothesis-only bias in natural language inference via ensemble adversarial training.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

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

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2021. Infobert: Improving robustness of language models from an information theoretic perspective. In *International Conference on Learning Representations*.

Dilin Wang, Chengyue Gong, and Qiang Liu. 2019a. Improving neural language modeling via adversarial training. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6555–6565. PMLR.

Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. 2019b. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 10506–10518.

Haohan Wang, Zexue He, Zachary C. Lipton, and Eric P. Xing. 2019c. Learning robust representations by projecting superficial statistics out. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14024–14031.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

Appendix

A Prompt templates

Inspired by Unified Prompts (Sanh et al., 2021) and Prompt Source project (Bach et al., 2022), we express all explanation prompting input-output pairs in the `jinj2` template language.[5] This choice allows us to take advantage of the many features and benefits offered by `jinja2`.

A.1 Explanation prompting template

Input:

```
Is this true and why?

    {{premise}} implies {{hypothesis}}
```

Output:

```
{% if label == 'entailment' %} Yes {%else%} No
{%endif%} it is  {{label}} because
{{explanation}}
```

A.2 Alternative template for training samples lacking explanations

Input:

```
Is this true?

    {{premise}} implies {{hypothesis}}
```

Output:

```
{% if label == 'entailment' %} Yes {%else%} No
{%endif%} it is {{label}}
```

Note that ANLI provides explanations for only some of the questions; where missing, the prompt template was modified to accommodate this.

B Training details

We fine-tuned the models on an Nvidia A100 node with 8 x 40GB GPUs. We used the DeepSpeed library[6] that implements ZeRo (Rajbhandari et al., 2020) and ZeRo-Offload (Ren et al., 2021); and the Huggingface transformers library (Wolf et al., 2019). We used an Adam optimizer (Kingma and Ba, 2015) with a learning rate of {1e-4, 5e-5}, with a per device batch size of {8, 16, 32, 64}, warm-up ratio of 0.08, max source length of 1024 except for T0 which uses 512 tokens with dynamic padding based on the longest sequence in the batch. We fine-tuned for a maximum of three epochs and selected the best checkpoint based on performance on the e-SNLI development set. Table 7 shows an overview of all used hyperparameters.

Models	
Warmup Ratio	0.08
Per Device Batch Size	{2, 4, 8, 16, 32, 64}
Learning Rate	{1e-3, 1e-4*, 1e-5, 5e-5*}
Adam ϵ	$1.00e-08$
Adam β_1	0.9
Adam β_2	0.999
Gradient Norm	1
Max Source Len	{512, 1024}
Max Target Len	256
weight_decay	0
fp16	yes
DeepSpeed	
fp16	
enabled	yes
loss_scale	0
loss_scale_window	1,000
initial_scale_power	16
hysteresis	2
min_loss_scale	1
zero_optimization	
sub_group_size	$1.00e9$
stage3_max_live_parameters	$1.00e9$
stage3_max_reuse_distance	$1.00e9$

Table 7: Hyperparameter settings. Where multiple values were tried the final values used is shown with an asterisk. The batch size of 8 is only used for the 11Billion parameter model.

C Prompt Sensitivity Results

We present extended results on prompt sensitivity with a range of model sizes, with all values being averages over three random seeds. Table 8 shows the results on each prompt. The aim of the experiment is to examine the sensitivity of various models to prompt wording. To do this, we evaluated the models on five diverse crowdsourced prompts obtained from the Prompt Source project (Bach et al., 2022). We conducted three separate experiments for each model, using three different random seeds, and report the average accuracy across all five prompts on the development sets of e-SNLI and Counterfactually-Augmented NLI. We limited the number of instances used from e-SNLI to twenty thousand randomly selected examples. The results demonstrate that models trained with explanation prompting outperform those trained with label-only across all five prompts in terms of accuracy.

D Crowd sourcing Forms

In this section, we present the crowdsourcing form utilized for the human evaluation of explanation

[5] https://https://jinja.palletsprojects.com/

[6] https://github.com/microsoft/DeepSpeed

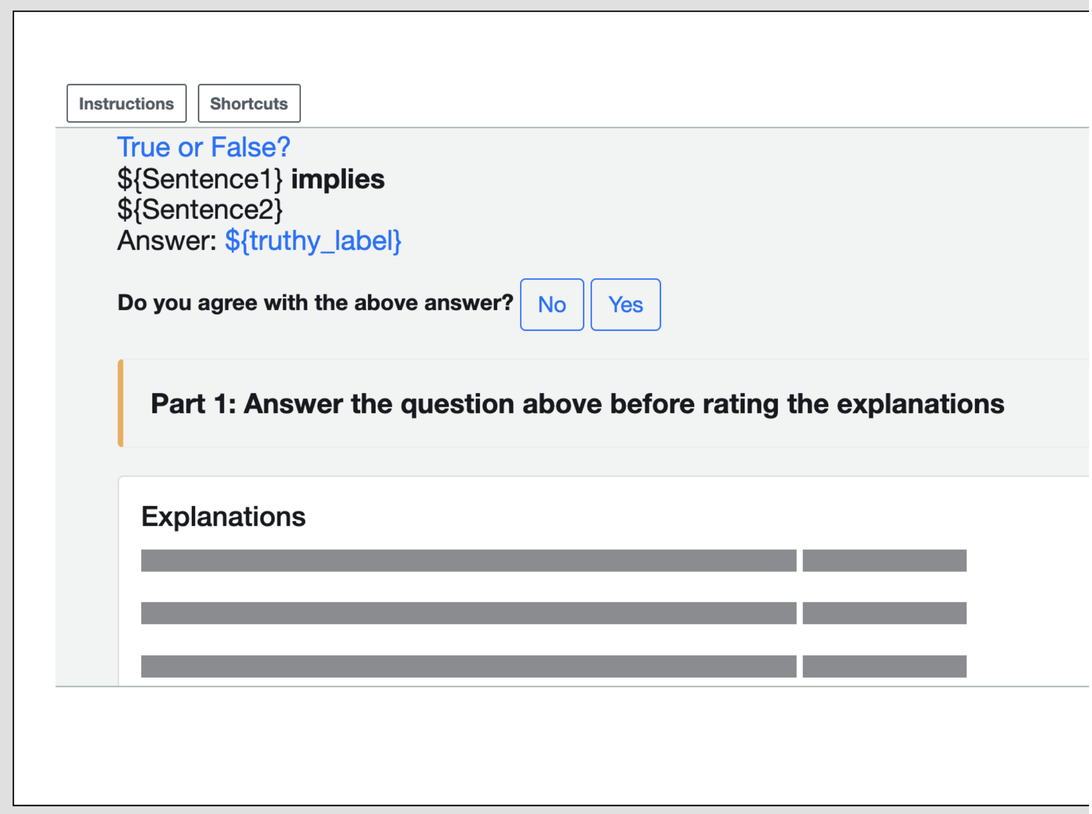
Dataset	Prompt_ID	T5-Small		T5-Large		T5-3B		BART-B		BART-L	
		LO	EP	LO	EP	LO	EP	LO	EP	LO	EP
e-SNLI	1	**71.1**	67.3	89.2	**93.3**	91.5	**94.4**	84.5	**90.1**	88.6	**93.0**
	2	66.7	**69.3**	89.6	**93.3**	91.6	**94.7**	84.3	**90.3**	59.1	**93.4**
	3	**72.3**	67.7	89.5	**93.2**	91.8	**94.6**	84.0	**90.5**	88.2	**93.3**
	4	**69.0**	66.4	89.6	**93.6**	91.6	**94.5**	84.1	**90.5**	72.3	**93.4**
	5	**69.2**	67.4	89.5	**93.4**	91.8	**94.4**	84.7	**90.3**	85.8	**93.3**
CNLI (RP)	1	44.8	**60.3**	64.4	**79.6**	70.8	**82.5**	50.1	**72.3**	60.1	**77.9**
	2	42.8	**68.5**	65.4	**79.2**	73.0	**83.8**	49.3	**71.8**	41.9	**77.2**
	3	47.5	**68.4**	65.9	**79.0**	71.5	**83.0**	49.3	**72.0**	59.9	**76.6**
	4	43.3	**68.3**	66.2	**79.3**	72.1	**83.0**	49.0	**71.8**	49.0	**79.0**
	5	43.4	**68.8**	65.9	**79.7**	70.1	**82.8**	48.8	**71.6**	58.3	**77.2**
CNLI (RH)	1	**65.6**	62.7	80.6	**91.1**	82.0	**92.3**	73.0	**86.5**	79.3	**90.8**
	2	54.8	**70.6**	81.1	**91.8**	82.8	**93.0**	72.2	**86.2**	54.1	**90.5**
	3	67.5	**69.4**	81.3	**91.3**	81.9	**92.8**	73.3	**85.7**	79.4	**91.0**
	4	63.2	**68.8**	81.6	**91.4**	82.0	**92.2**	73.0	**87.3**	65.6	**90.8**
	5	60.7	**68.7**	82.3	**91.1**	83.1	**92.5**	71.8	**85.3**	77.0	**90.9**
CNLI (RP&RH)	1	55.2	**61.5**	72.5	**85.3**	76.4	**87.4**	61.5	**79.4**	69.7	**84.3**
	2	48.8	**69.5**	73.3	**85.5**	77.9	**88.4**	60.7	**79.0**	48.0	**83.8**
	3	57.5	**68.9**	73.6	**85.1**	76.7	**87.9**	61.3	**78.8**	69.7	**83.8**
	4	53.3	**68.5**	73.9	**85.3**	77.0	**87.6**	61.0	**79.5**	57.3	**84.9**
	5	52.0	**68.8**	74.1	**85.4**	76.6	**87.7**	60.3	**78.4**	67.7	**84.0**

Table 8: Prompt-sensitivity results on the development set of e-SNLI and Counterfactually-Augmented NLI (CNLI). The values represent mean accuracy over three random seeds. The table compares the accuracy of a label-only prompting model (represented by the LO column) and the explanation prompting model (represented by the EP column). Explanation prompting models outperform the label-onlymodel on almost all the prompts.

quality. To ensure unbiased assessments, the explanations were initially concealed from the crowd workers while they read the question and answer (as depicted in Fig.5a). Once the workers answered the first question, the explanations were then revealed (as shown in Fig.5b).

E Qualitative Results

Table 9 shows some qualitative examples of the *explanations* generated by the explanation prompting T0 model.



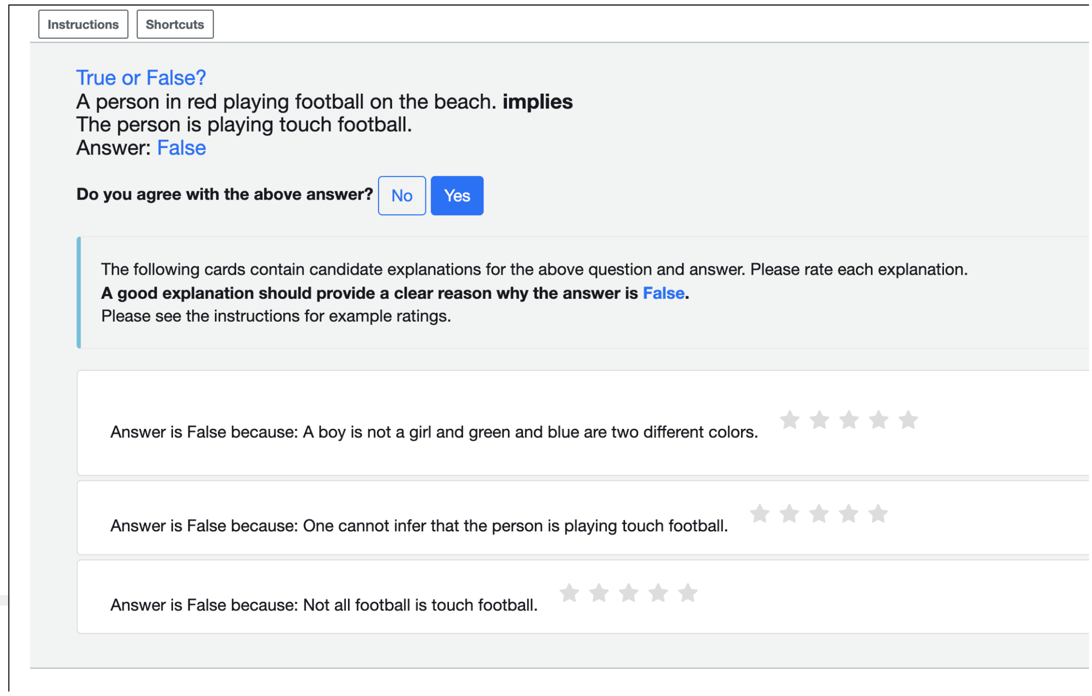
True or False?
${Sentence1} **implies**
${Sentence2}
Answer: ${truthy_label}

Do you agree with the above answer?

Part 1: Answer the question above before rating the explanations

Explanations

(a) The figure depicts the process of evaluating the quality of the explanations in the crowd-sourcing task. The first stage requires the user to carefully read both the premise and hypothesis presented. After familiarizing themselves with the information, the user is then tasked with answering the corresponding question. At this stage, the explanations for the question are initially hidden.



True or False?
A person in red playing football on the beach. **implies**
The person is playing touch football.
Answer: False

Do you agree with the above answer?

The following cards contain candidate explanations for the above question and answer. Please rate each explanation.
A good explanation should provide a clear reason why the answer is False.
Please see the instructions for example ratings.

Answer is False because: A boy is not a girl and green and blue are two different colors.

Answer is False because: One cannot infer that the person is playing touch football.

Answer is False because: Not all football is touch football.

(b) The figure illustrates the stage of the crowdsourcing process where the explanations are reviewed after the first question has been answered (figure a). Hovering over the rating stars reveals the meaning of the star such as "Very bad" for 1 star and "Bad" for 2 stars. Upon choosing a star rating, the form displays its corresponding description, such as "Very bad" for 1 star, "Bad" for 2 stars, "Good" for 3 stars, "Very good" for 4 stars, and "Extremely good" for 5 stars.

#	Input	Reason	Target
1	**P**: Two women are embracing while holding to go packages. **H:** The sisters are hugging goodbye while holding to go packages after just eating lunch.	**Model**: Two women are not necessarily sisters. **Human:** The to go packages may not be from lunch.	N (o)
2	**P**: Two women are embracing while holding to go packages. **H:** Two woman are holding packages.	**Model**: Two women are holding packages is the same as two women are holding packages. **Human:** Saying the two women are holding packages is a way to paraphrase that the packages they are holding are to go packages.	E (o)
3	**P**: Two women are embracing while holding to go packages. **H:** The men are fighting outside a deli.	**Model**: Women are not men. **Human:** In the first sentence there is an action of affection between women while on the second sentence there is a fight between men.	C (o)
4	**P**: Two young children in blue jerseys, one with the number 9 and one with the number 2 are standing on wooden steps in a bathroom and washing their hands in a sink. **H:** Two kids at a ballgame wash their hands.	**Model**: Bathroom and ballgame are different. **Human:** Two kids in jerseys watching their hands are not necessarily at a ballgame.	N (x)
5	**P**: A woman is doing a cartwheel while wearing a bikini in the sand next to the beach. **H:** A woman is doing a cartwheel and falls on her head.	**Model**: Just because a woman is doing a cartwheel doesn't mean she falls on her head. **Human:** A woman doing a cartwheel as opposed to a woman falling while doing a cartwheel.	C (x)
6	**P**: At an outdoor event in an Asian-themed area, a crowd congregates as one person in a yellow Chinese dragon costume confronts the camera. **H:** A single man is next to a camera	**Model**: A single man is a person. **Human:** The crowd can't be single man.	C (x)

Table 9: Examples explanations generated by our models on the development set of e-SNLI, which consists of three labels: entailment (E), contradiction (C) and neutral (N). We show the first three correct and incorrect instances.