To what extent do human explanations of model behavior align with actual model behavior?

Grusha Prasad[†], **Yixin Nie**[‡], **Mohit Bansal**[‡], **Robin Jia**[⋆], **Douwe Kiela**[⋆], **Adina Williams**[⋆]
[†] Johns Hopkins University; [‡] UNC Chapel Hill; [⋆] Facebook AI Research
grusha.prasad@jhu.edu, adinawilliams@fb.com

Abstract

Given the increasingly prominent role NLP models (will) play in our lives, it is important to evaluate models on their alignment with human expectations of how models behave. Using Natural Language Inference (NLI) as a case study, we investigated the extent to which human-generated explanations of models' inference decisions align with how models actually make these decisions. More specifically, we defined two alignment metrics that quantify how well natural language human explanations align with model sensitivity to input words, as measured by integrated gradients. Then, we evaluated six different transformer models (the base and large versions of BERT, RoBERTa and ELECTRA), and found that the BERT-base model has the highest alignment with human-generated explanations, for both alignment metrics. Additionally, the base versions of the models we surveyed tended to have higher alignment with human-generated explanations than their larger counterparts, suggesting that increasing the number model parameters could result in *worse* alignment with human explanations. Finally, we find that a model's alignment with human explanations is not predicted by the model's accuracy on NLI, suggesting that accuracy and alignment are orthogonal, and both are important ways to evaluate models.

1 Introduction

NLP models often make classification decisions in ways humans don't expect. For example, QA models often choose the correct answer for one example, but fail catastrophically on very similar examples (Ribeiro et al., 2018; Wallace et al., 2019; Selvaraju et al., 2020), such as answering "Is the rose red?" with no, but then "What color is the rose?" with "red" (Ribeiro et al., 2019). VQA models often attend to different portions of images than humans do (Das et al., 2016). NLI models often over-attend to particular words or rely on shallow heuristics, and can often perform unexpectedly well from only looking at the hypothesis (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018). Since people generally do not expect models to base decisions on spurious correlations in the data (cf. McCoy et al. 2019), models that align with human expectations are less likely to make the right decisions for the wrong reasons.

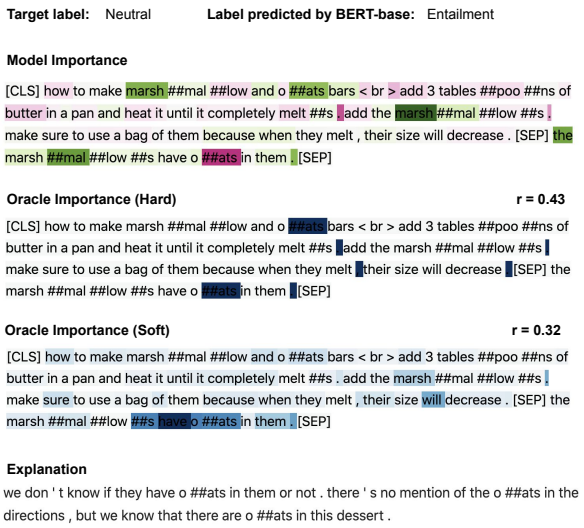

Figure 1: An example illustrating different token-level importance values. "Model Importance" is color coded by integrated gradients attribution for BERT-base (red indicates negative and green indicates positive attribution). The other two rows show the oracle importance scores estimated by the hard and soft oracles (darker values indicate more important).

In this paper, we measure how well model decisions are aligned with human expectations about those decisions. Building on work that aims to extract or generate interpretable or faithful descriptions of model behavior (Lipton, 2018; Rajani et al., 2019; Kalouli et al., 2020; Silva et al., 2020; Jacovi and Goldberg, 2020; Zhao and Vydiswaran, 2020), we use human-generated natural language explanations to determine which portions of the input

people expected to be *important* for the models' decisions. We then use Integrated Gradients (IG, Sundararajan et al. 2017) to determine which portions of the input actually influenced the models' decisions. We term the alignment between the two *importance alignment*.

We approach the question in the context of the Natural Language Inference task (Dagan et al., 2006; Bowman et al., 2015; Williams et al., 2018) which requires models to classify examples according to whether they are entailing, contradicting, or neutral. Specifically, we use the human-generated explanations in the Adversarial NLI dataset (ANLI, Nie et al. 2020) as our ground truth measure for two reasons. First, the explanations in the ANLI dataset were collected in an adversarial setting. Annotators were given a context and a label and were asked to write a hypothesis that fooled a target model; if the model was fooled, annotators explained in natural language why they thought they were successful. Such an adversarial setting encourages annotators to reason about the models' decision making process. Second, the explanations include information not only about why example should receive the gold label, making this dataset more suitable for our purposes than other explanation datasets such as e-SNLI (Camburu et al., 2018). The ANLI dataset has the added benefits of having been collected in multiple genres over several rounds, and having example-level hand-annotations (Williams et al., 2020).

Our contributions are as follows. We formulate two measures of importance alignment based on two different methods for converting natural language explanations to the parts of the input people expected to be *important* in influencing model decisions. Then, we evaluate six state-of-the-art transformer NLI models and find that according to both measures, BERT-base has a much higher importance alignment score than any of the other models. Additionally, the smaller versions of a model (i.e., the base versions) tended to have higher importance alignment than the corresponding larger versions. Finally, we demonstrate that more accurate models do not necessarily have higher importance alignment, suggesting that accuracy and alignment with human expectations are orthogonal dimensions along which models should be evaluated.

2 Related Work

To some AI practitioners, "alignment" evokes thoughts of "value alignment", namely the idea that models should be aligned to normative notions of human ethics (Russell et al., 2015; Peng et al., 2020). To a specialist in machine translation, "alignment" refers to mapping tokens from source to target, while in multimodal circles, "alignment" is the correspondence between images and text (e.g., an image-caption alignment model), etc. In this paper, we suggest a new type of alignment, namely importance alignment: not only do we want high performing models, we also want models that behave in ways that people expect them to. We want models that are *aligned* to human expectations (and vice versa).

Although high alignment in some contexts can cause problems—for example, leading us to inappropriately prefer decisions made by AI systems to ones made by humans (Araujo et al., 2020) or to trust AI recommendations too much (Wagner et al., 2018; Howard, 2020; Smith-Renner et al., 2020), there are many reasons to want aligned models. For example, providing annotators with additional information about model decisions, say model accuracy (cf. Yin et al. 2019), or an generated explanation (Bansal et al., 2020b) can cause their level of trust to rise, and in some settings, can actually decrease mistakes (Zhang et al., 2020). Similarly, when humans have a good theory of "mind" for a model, debugging might be easier, since errors will be easier to spot. Humans who have formed a correct mental model of AI decision boundaries make better AI-assisted decisions (Bansal et al., 2019a,b). In the words of Bansal et al. (2020a), "predictable performance is worth a slight sacrifice in AI accuracy", especially on tasks with potentially serious social implications, making tight importance alignment a worthy goal.

3 Methods

We compute importance alignment ($\mathcal{A}$) between two types of importance scores ($\mathbb{I}$): first, **model importance** ($\mathbb{I}_m(x, y)$) which quantifies the extent to which different parts of the input influenced the models' decisions; second, **oracle importance** ($\mathbb{I}_m^O(x, y)$) which quantifies the extent to which annotators expected the different parts of the input to influence the models' decisions.[1]

[1] We refer to this as the oracle, because we consider importance scores derived from human-generated explanations to

3.1 Calculating Model Importance

We measure model importance using integrated gradients because they are axiomatically both interpretable and faithful (Sundararajan et al., 2017)[2], thereby avoiding some of the pitfalls of attention Jain and Wallace (2019). Concretely, for a model m, model importance is defined for each token in an example x—a three-tuple of context x^c, a hypothesis x^h and (optionally) and explanation x^e—with gold label y as follows:

$$\mathbb{I}_m(x, y) = |IG_m(x^c, x^h, y)| \quad (1)$$

where IG_m returns a vector of IG attribution scores for each token in the concatenation of x^c and x^h, and $|\cdot|$ denotes component-wise absolute value. We used absolute value of IG because ANLI annotators were instructed to provide an explanation for both (i) why the annotated label is correct, and (ii) why they think this example was difficult for the system to label. If annotators provided explanations about the correct behaviour (i.e., why the gold label was correct), then annotator explanations should correlate with positive attributions. On the other hand, if annotators provided explanations about incorrect behaviour (i.e., why the model was wrong), then annotator explanations should correlate with negative attributions. Therefore by taking the absolute value, we are agnostic to the explanation type and instead identify tokens in the input that strongly influence the model's prediction (either towards or away from the gold label).

We note that there exist other options to calculate model importance such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017) and their variants, although such perturbation-based explanability methods have been argued to be unreliable (Slack et al., 2020). No current consensus exists on which methods should be employed (Hase and Bansal, 2020). Although we use IG, crucially, our method is not dependent on it; IG can be replaced with any method that can faithfully assign an importance score to each token in the input.

3.2 Calculating Oracle Importance

There is no consensus on how best to convert natural language explanations into importance scores indicating how much annotators expect input tokens to influence models' decisions. Therefore, we define two new measures of oracle importance: hard oracle importance ($\mathbb{I}^H_m(x, y)$) and soft oracle importance ($\mathbb{I}^S_m(x, y)$).

Hard Oracle Importance. Hard oracle importance is a binary measure which uses token overlap between the context and hypothesis taken together and the explanation. Intuitively, when annotators thought a token in the context or hypothesis was important, they would likely use that token in their explanation. Formally, this can be expressed as

$$\mathbb{I}^H_m(x, y) = overlap_m(x^c, x^h, x^e), \quad (2)$$

where the overlap function yields a binary vector where the i-th component is valued 1 if $t_i \in x^e$ for context and hypothesis tokens t or else 0.

By and large, the explanations provided by the annotators are model agnostic, reflecting expectations of model behavior in general, not the behavior of some specific model. However, to enable apples-to-apples comparison between the hard oracle and model importances, we must operate on the same precise tokens. Therefore, $\mathbb{I}^H_m(x, y)$ is model specific to the extent that we use model specific tokenizers.

Soft Oracle Importance. The hard oracle is very simple and does not capture synonyms, entities referenced by pronouns, syntactic cues, etc. To overcome these shortcomings, we also define soft oracle importance where we compute importance scores from IG of explanation-informed models. Concretely, this can be expressed as:

$$\mathbb{I}^S_m(x, y) = |IG_{m'}(x^c, x^h, x^e, y)|. \quad (3)$$

To calculate $\mathbb{I}^S_m(x, y)$, we constructed an artificial classification task that incorporates the human-written explanations: given a context x^c, hypothesis x^h, and explanation x^e, jointly predict both the entailment relationship between x^c and x^h as well as whether x^e was written about (x^c, x^h) or about a different example. In this task, predicting the gold label requires the model to not only perform inference, but also to establish a relationship between the provided explanation and the input, thereby incorporating information from the natural language explanation. For each of our target model architectures, we fine-tune three classifiers corresponding to m' in the above equation (with different random seeds) to perform this as a 6-way classification task. Accuracy of these models is provided in the Appendix in Table 2.

be the ground truth.

[2] https://captum.ai/

We generated the training ($n = 19043$) and development datasets ($n = 2116$) for these classifiers by subsetting the portion of the ANLI training set for which a explanation was provided. Recall that only model-fooling examples were provided with explanations in ANLI. Each original premise-hypothesis pair in our training dataset appeared twice: once with a matched explanationand once with a randomly selected explanation. We fine-tuned the models for two epochs. The mean accuracy of these classifiers across seeds for each architecture on the ANLI development set is provided in the supplementary materials.

Why not directly collect oracle importance scores from annotators? We convert existing natural language explanations into importance scores instead of directly collecting importance scores from annotators for two reasons. First, we think that for most non-expert annotators, asking them to provide verbal descriptions is easier and more natural than asking them to answer a question like "For which words do you think the model's prediction would change the most if that word was blanked out?". Furthermore, if one wanted to pursue this angle assiduously, they would need annotators who know what IG is and ask them to predict IG scores—after all, if we're defining "how the model works" as IG scores, then "how do humans think models works" should be defined as "what do humans think IG scores will be". Of course, asking annotators to do this is highly impractical. Second, the ANLI dataset already makes the natural language data available, and for the reasons described earlier, the adversarial context in which the ANLI data was collected makes it particularly suitable for our purposes.

3.3 Calculating Importance Alignment

We calculate importance alignment as the mean product-moment (i.e., Pearson's) correlation between model importances and oracle importances (soft or hard) at the token level as

$$C_m(x, y) = r(\mathbb{I}_m(x, y), \mathbb{I}^O_m(x, y)), \qquad (4)$$

where the oracle O is either the hard (H) or the soft (S) oracle. Since taking the mean of non-transformed correlation values can result in biased estimates (Fisher, 1921), we Fisher z-transform the correlation C, and take the mean over all examples in our dataset D as:

$$\mathcal{A} = \tanh\left(\frac{1}{\mid D \mid}\sum_{(x,y)\in D}\operatorname{arctanh}(C_m(x, y))\right) \qquad (5)$$

to derive **hard** ($\mathcal{A}^H$) and **soft oracle alignment** ($\mathcal{A}^S$) respectively.

Since the explanations were provided only when the model was fooled, $\mathcal{A}$ is calculated only over the examples that the target models got wrong.

A Baseline: Random Explanations. Consider a scenario where the second word of the input always received a high attribution value for BERT models. In this case, we might expect a non-zero correlation between model importance and soft oracle importance for BERT, even if the model behaviour was not aligned with human explanations at all. Similarly, we might also expect non-zero correlation if specific words received a high attribution no matter what context they occur in, and these words occur frequently in both the input and the explanations. Given D, we construct a dataset D' by randomly associating each context-hypothesis pair with an unrelated explanation. We define the random baseline alignment score $\mathcal{A}_R$ as the importance alignment whose oracle alignment was computed on D' instead of D. Given that $\mathcal{A}_R$ can differ across model architecture, in all of our analyses we consider $\Delta\mathcal{A}$, which can be given as follows:

$$\Delta\mathcal{A} = \mathcal{A} - \mathcal{A}_R \qquad (6)$$

3.4 Target models

We calculated $\mathcal{A}$ and $\mathcal{A}_R$ for 6 Transformer-based models pretrained on a language modeling objective, including BERT base and large (Devlin et al., 2019); RoBERTa base and large (Liu et al., 2019); and ELECTRA base and large (Clark et al., 2020). All models used to calculate $\mathbb{I}_m(x, y)$ were trained on a combination of SNLI, MultiNLI and NLI-recast FEVER (Thorne et al., 2019) and all rounds of ANLI. We fine-tuned five different models for each architecture with different random seeds.[3]

4 Results

In Table 1, we report importance alignment for both oracles for all tested models (averaged across 5 seeds).

[3] All the models that we calculate model importance from, as well as the code use to calculate the model importance and run the statistical analyses will be available upon publication.

Model	Importance Alignment				Acc.
	$\mathcal{A}^H$	$\Delta\mathcal{A}^H$	$\mathcal{A}^S$	$\Delta\mathcal{A}^S$	ANLI
BERT-Base	0.23	0.21^{***}	0.33	0.11^{***}	48.02
RoBERTa-Base	0.21	0.11^{***}	0.29	0.02^{*}	50.47
ELECTRA-Base	0.17	0.17^{***}	0.26	0.06^{***}	52.33
BERT-Large	0.15	0.18^{***}	0.15	-0.02	49.24
RoBERTa-Large	0.21	0.04^{*}	0.23	0.01	55.37
ELECTRA-Large	0.15	0.07	0.11	0.01	58.06
All Base models	0.21	0.17	0.30	0.07	50.27
All Large models	0.17	0.11	0.16	-0.00	54.56

Table 1: Importance Alignment between model importance scores and oracle importance scores (both $\mathcal{A}^H$ and $\mathcal{A}^S$ metrics) across 5 random seeds on the ANLI dataset. $\mathcal{A}$ was computed only over the examples that the models got wrong. Average model accuracy across seeds and different rounds of ANLI is also provided. '*'s indicate whether $\Delta\mathcal{A}^H$ and $\Delta\mathcal{A}^S$ are significant based on a paired t-test between $\mathcal{A}$ and $\mathcal{A}_R$. '***' indicates $p < 0.001$, '**' indicates $p < 0.01$ and '*' indicates $p < 0.05$.

For each of our six models, we calculated whether importance alignment with original examples was significantly greater than importance alignment for random examples for both measures of Oracle importance (i.e., whether $\mathcal{A}^H > \mathcal{A}^H_R$ and whether $\mathcal{A}^S > \mathcal{A}^S_R$). Specifically, for each example, we estimated the correlation between model importance ($\mathbb{I}_m(x, y)$) and oracle importance ($\mathbb{I}^O_m(x, y)$) as described in Equation 4 for both the original explanations and the random ones. Then, we Fisher-transformed these example-level correlation values and used paired t-tests to determine whether the Fisher-transformed correlation for the original examples was significantly greater than the Fisher-transformed correlation for the random examples across all random seeds. Statistical significance is indicated in Table 1 (untransformed to be more interpretable as r-correlation coefficients) with asterisks (*). See Figure 3 in the Appendix for by-seed results.

For both types of oracles, $\Delta\mathcal{A}$ was significant for all of the base versions of the models, with BERT-base having the highest value of $\Delta\mathcal{A}$. For the large versions of BERT and RoBERTa, alignment measured using the hard oracle ($\Delta\mathcal{A}^H$) was significant, but alignment measured using the soft oracle ($\Delta\mathcal{A}^S$) was not. The magnitude of $\Delta\mathcal{A}^H$ was much larger for BERT-large than for RoBERTa-large. Finally, neither $\Delta\mathcal{A}^H$ nor $\Delta\mathcal{A}^S$ were significant for ELECTRA-large.

Taken together, these results indicate the following about the alignment between how models make inference decisions and how human expect them to: base models have stronger alignment than do large models, with BERT-base having the strongest alignment of all. However, it is important to note that even with the BERT-base models, the magnitude of $\Delta\mathcal{A}$ is moderately small, suggesting that there is ample room for improvement.

Accuracy does not predict importance alignment. Since the base models have a greater importance alignment but lower accuracy than their larger counterparts, it seems plausible that accuracy is negatively correlated with importance alignment. To test this, we computed a separate value of $\Delta\mathcal{A}^H$ and $\Delta\mathcal{A}^S$ for each random seed of the target model and for the three different rounds of ANLI. Then, we computed the product moment (i.e., Pearson's) correlation between model accuracy on NLI and $\Delta\mathcal{A}^H$ as well as model accuracy and $\Delta\mathcal{A}^S$ and found no significant correlation for either the hard (-0.03, $p = 0.76$) or soft (-0.08, $p = 0.46$) oracles. We plot this relationship between importance alignment and accuracy in Figure 2.

$\mathcal{A}^H$ and $\mathcal{A}^S$ differ. While in almost all cases, $\Delta\mathcal{A}^H$ was greater than $\Delta\mathcal{A}^S$, we do not take this to mean that $\mathcal{A}^H$ is necessarily a better measure of importance alignment than $\mathcal{A}^S$, or that $\mathbb{I}^H_m(x, y)$ is necessarily a better measure of oracle importance than $\mathbb{I}^S_m(x, y)$. Clearly, we cannot use the alignment scores to evaluate our methods since we used the same methods to convert natural language explanations to oracle importance scores and calculate the alignment scores in the first place: for this, an external measure independent of alignment with model importance would be required.

While one could imagine running large scale annotation experiments to measure which of our importance metrics people think is "better", we think that picking a "best" conversion method is less ideal given the data than considering both in tandem. To illustrate our point, let us consider the example in Figure 1. The hard overlap measure fails to assign importance to words that it should assign importance to given the explanation (e.g., "marshmallow") whereas the soft overlap assigns higher importance to words that humans likely wouldn't think were important (e.g., "sure" and "will"). Thus, both methods are imperfect in different ways. Considering the importance alignment scores from both methods together provides a more

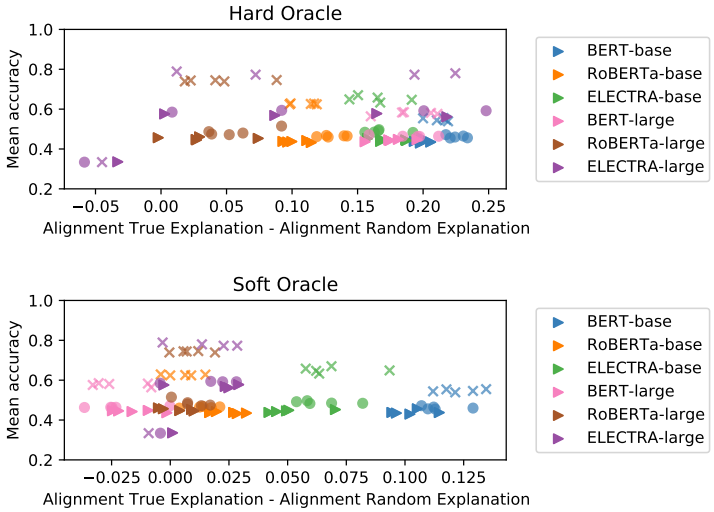

Figure 2: Accuracy and seed don't explain our significant correlations for either $\mathcal{A}^H$ ("overlap with explanation") or $\mathcal{A}^S$ ("NLI with explanation"). The cross, circle, and triangle refer to rounds 1, 2, and 3 of ANLI respectively.

holistic view of whether the models' decisions are aligned with human expectations. Therefore, the strongest conclusions about importance alignment that can be drawn hold for both $\mathcal{A}^H$ and $\mathcal{A}^S$.

5 Discussion & Conclusion

We evaluated six models on two measures of importance alignment, and found that for both measures, the importance alignment scores for the the base versions of the models was greater than for their larger counterparts, with the BERT-base models having the highest importance alignment. This suggests that the NLI decision making process of models with more parameters is *less* aligned with human expectations of how the models (should) make decisions when compared to models with fewer parameters.

Future work. While we used Integrated Gradients to measure model importance, conceptually, our proposed method is agnostic to how we calculate model importance. Future work could explore other existing methods to calculate model importance scores (e.g., LIME, SHAP, and their variants). Similarly, we defined two methods of converting natural language explanations to oracle importance scores. Although we think these methods are reasonable starting points, further work is required both to validate these methods as well as explore other conversion methods such as using explanation generation (e.g., Hase et al. 2020; Camburu et al. 2018). Future work can also explore the consequences of using different types of explanation to measure importance alignment (e.g., separating explanations for why the model is wrong from why the label is correct, using human importance annotations instead of oracle importance measures derived from natural language explanations etc.).

Further research could verify and build on our results, by asking, for example, do non-transformer models with fewer parameters also have higher alignment than other non-transformer models with more features? Is there a threshold for model parameters, where the inverse relationship between model size and alignment score breaks down for models with parameters less than this threshold? Is there a similar threshold for model accuracy? How do other factors like number of training examples, types of training examples and training objective impact importance alignment? The answers to these questions can not only result in models better aligned to human expectations, but can also provide insight into what factors are important for better alignment with human expectations.

Summary. As our reliance on NLP models increases, it is not only important that these models make accurate decisions, but also that the decisions they make align with humans expectations. In this paper, we introduced importance alignment as a metric to measure the alignment between model decisions and people's explanations of these decisions. We calculated importance alignment by estimating two types of token level importance scores: model importance score (to what extent did a token influence the models' decision) and oracle impor-

tance score (to what extent did people expect a token to influence the models' decision). While we were able to use existing methods from the interpretability literature to faithfully estimate the model importance score, no such method exists to convert natural language explanations to oracle scores. Therefore, we proposed two methods, resulting in two types of oracle scores, hard and soft. From these two types of oracle scores, we calculated two versions of importance alignment: $\mathcal{A}^H$ and $\mathcal{A}^S$.

For all of the models we surveyed, there was a greater importance alignment to matched explanations than to random explanations, suggesting human expectations aligned to some extent, although correlations were weak. We also found that importance alignment was not predicted by accuracy: more accurate models did not necessarily have better importance alignment. This suggests that accuracy and importance alignment are potentially orthogonal dimensions along which models can be evaluated. Importance alignment was however predicted by model size: there was a greater importance alignment for smaller versions of a given model architecture type that for the larger versions. This suggests that models with fewer parameters might be more aligned with human expectations.

Acknowledgements

Thank you to Luke Zettlemoyer, Ana Valeria Gonzalez, the Dynabench team, Roy Schwartz, and to Tal Linzen's Computation and Psycholinguistics lab for invaluable comments and support!

References

Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, pages 1–13.

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2020a. Optimizing AI for teamwork. *arXiv preprint arXiv:2004.13102*.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019a. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019b. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Túlio Ribeiro, and Daniel S. Weld. 2020b. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. *arXiV*, abs/2006.14779.

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.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31, pages 9539–9549. Curran Associates, Inc.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, Austin, Texas. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ronald Aylmer Fisher. 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. *Biometrika*, 10(4):507–521.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. 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.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4351–4367, Online. Association for Computational Linguistics.

Ayanna Howard. 2020. Are we trusting AI too much? Examining human-robot interactions in the real world. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1–1.

Alon Jacovi and Yoav Goldberg. 2020. Aligning faithful interpretations with their social attribution. *arXiV*, abs/2006.01067.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Valeria de Paiva, Richard Crouch, and Mennatallah El-Assady. 2020. XplaiNLI: Explainable natural language inference through visual analytics. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 48–52.

Zachary C Lipton. 2018. The mythos of model interpretability. *Queue*, 16(3):31–57.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

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.

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*, pages 4885–4901, Online. Association for Computational Linguistics.

Xiangyu Peng, Siyan Li, Spencer Frazier, and Mark Riedl. 2020. Reducing non-normative text generation from language models. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 374–383, Dublin, Ireland. 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.

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

Marco Túlio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6174–6184. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 856–865. Association for Computational Linguistics.

Stuart Russell, Daniel Dewey, and Max Tegmark. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4):105–114.

Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Túlio Ribeiro, Besmira Nushi, and Ece Kamar. 2020. Squinting at VQA models: Introspecting VQA models with sub-questions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10000–10008. IEEE.

Vivian S Silva, André Freitas, and Siegfried Handschuh. 2020. XTE: Explainable text entailment. *arXiv preprint arXiv:2009.12431*.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, pages 180–186. ACM.

Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan L. Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ML. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alan R Wagner, Jason Borenstein, and Ayanna Howard. 2018. Overtrust in the robotic age. In *Communications of the Association for Computing Machinery*, volume 61, pages 22–24. ACM New York, NY, USA.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2020. Anlizing the adversarial natural language inference dataset.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305.

Xinyan Zhao and V. G. Vinod Vydiswaran. 2020. LIREx: Augmenting language inference with relevant explanation. *arXiv preprint arXiv:2012.09157*.

	Target models	Soft oracle
BERT-Base	48.02	44.85
RoBERTa-Base	50.47	60.93
ELECTRA-Base	52.33	51.93
BERT-Large	49.24	49.01
RoBERTa-Large	55.37	74.21
ELECTRA-Large	58.06	74.93

Table 2: Accuracy on the development partition of the ANLI dataset for target models (finetuned on NLI with MNLI + SNLI + ANLI + FEVER re-cast as NLI) and models used as the soft oracle (finetuned on 6-way NLI and reason classification on a subset of the ANLI dataset). The accuracy for the target models is averaged over 5 random seeds and the accuracy for the soft oracle models is averaged over 3 random seeds. Even though the soft oracle models were trained on a smaller dataset, and the accuracy reflects a 6-way classification accuracy in comparison to the 3-way classification accuracy of the NLI models, the accuracy of the soft oracle models are equivalent to and in many cases even greater than the accuracy of the NLI models. This suggests that the soft oracle models did incorporate information about explanations, and that the information present in the explanations was useful for NLI.

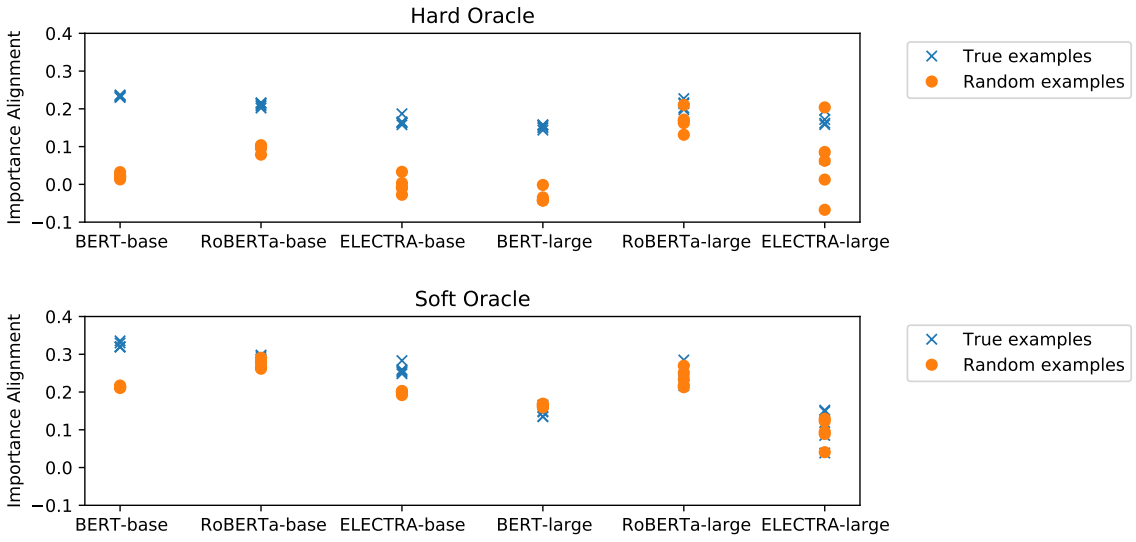

Figure 3: Alignment scores for $\mathcal{A}^H$ vs. $\mathcal{A}^H_R$ (top panel) and $\mathcal{A}^S$ vs. $\mathcal{A}^S_R$ (bottom panel) for 5 seeds for each of the models we surveyed. $\mathcal{A}^H$ and $\mathcal{A}^S$ are represented with blue crosses and $\mathcal{A}^H$ and $\mathcal{A}^S_R$ are represented with orange triangles. All the models, except ELECTRA-large have very low variability in the importance alignment scores across seeds.