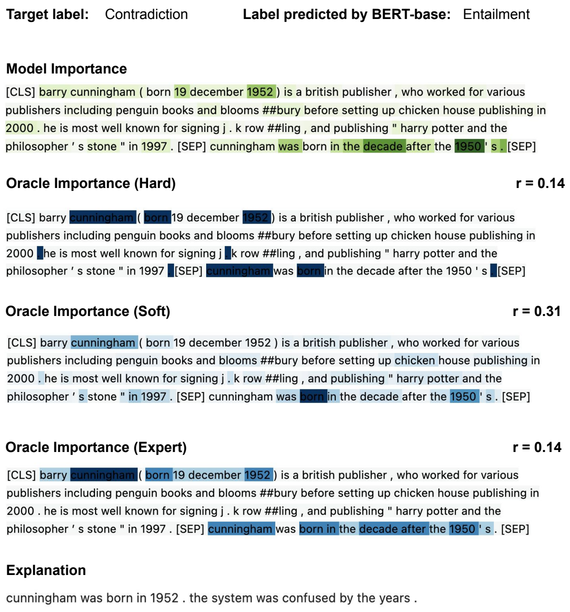
To what extent do human explanations of model behavior align with actual model behavior?

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 for human expectations of model behavior to align with actual model behavior. Using Natural Language Inference (NLI) as a case study, we investigate the extent to which human-generated explanations of models' inference decisions align with how models actually make these decisions. More specifically, we define three alignment metrics that quantify how well natural language explanations align with model sensitivity to input words, as measured by integrated gradients. Then, we evaluate eight different models (the base and large versions of BERT, RoBERTa and ELECTRA, as well as an RNN and bag-of-words model), and find that the BERT-base model has the highest alignment with human-generated explanations, for all alignment metrics. Focusing in on transformers, we find that the base versions tend to have higher alignment with human-generated explanations than their larger counterparts, suggesting that increasing the number of model parameters leads, in some cases, to *worse* alignment with human explanations. Finally, we find that a model's alignment with human explanations is not predicted by the model's accuracy, suggesting that accuracy and alignment are complementary ways to evaluate models.



Target label: Contradiction **Label predicted by BERT-base:** Entailment

Model Importance

[CLS] barry cunningham (born 19 december 1952) is a british publisher , who worked for various publishers including penguin books and blooms ##bury before setting up chicken house publishing in 2000 . he is most well known for signing j . k row ##ling , and publishing " harry potter and the philosopher ’ s stone " in 1997 . [SEP] cunningham was born in the decade after the 1950 ' s . [SEP]

Oracle Importance (Hard) **r = 0.14**

[CLS] barry cunningham (born 19 december 1952) is a british publisher , who worked for various publishers including penguin books and blooms ##bury before setting up chicken house publishing in 2000 . he is most well known for signing j . k row ##ling , and publishing " harry potter and the philosopher ’ s stone " in 1997 . [SEP] cunningham was born in the decade after the 1950 ' s . [SEP]

Oracle Importance (Soft) **r = 0.31**

[CLS] barry cunningham (born 19 december 1952) is a british publisher , who worked for various publishers including penguin books and blooms ##bury before setting up chicken house publishing in 2000 . he is most well known for signing j . k row ##ling , and publishing " harry potter and the philosopher ’ s stone " in 1997 . [SEP] cunningham was born in the decade after the 1950 ' s . [SEP]

Oracle Importance (Expert) **r = 0.14**

[CLS] barry cunningham (born 19 december 1952) is a british publisher , who worked for various publishers including penguin books and blooms ##bury before setting up chicken house publishing in 2000 . he is most well known for signing j . k row ##ling , and publishing " harry potter and the philosopher ’ s stone " in 1997 . [SEP] cunningham was born in the decade after the 1950 ' s . [SEP]

Explanation

cunningham was born in 1952 . the system was confused by the years .

Figure 1: An example illustrating different token-level importance values. “Model Importance” is color coded by absolute value integrated gradients attribution for BERT-base. The other three rows show the oracle importance scores estimated by the hard, soft and expert oracles (darker values indicate more important).

1 Introduction

NLP models often make classification decisions in ways humans don't expect them to. For example, Question Answering (QA) models often choose the correct answer for one example, but fail catastrophically on other 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 rely on shallow heuristics, their predictions are inappropriately affected by particular words, and they sometimes 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 make decisions in alignment with human expectations are less likely to make the right decisions for the wrong reasons.

In this paper, we measure how well model decisions are aligned with human expectations

[*] The work was conducted during an internship at Facebook AI Research.

Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 1–14
Online, November 11, 2021. ©2021 Association for Computational Linguistics

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* in influencing models' decisions. We then use Integrated Gradients (IG, Sundararajan et al. 2017) to determine which portions actually influenced the models' decisions. We term the alignment between them as **Importance Alignment**. We formulate three different methods of using human-generated natural language explanations to quantify human expectations of model behavior, resulting in three different methods for calculating importance alignment.

As a case study, we applied our method to the Natural Language Inference task (Dagan et al., 2006; Bowman et al., 2015; Williams et al., 2018) in which models are tasked with classifying pairs of sentences according to whether the first sentence entails, contradicts, or is neutral with respect to the second. Concretely, we measured the extent to which the inference decisions of eight models (six state-of-the-art transformers, an LSTM model and a bag-of-words model) aligned with human-generated explanations from the Adversarial NLI dataset (ANLI, Nie et al. 2020).

In all three methods for calculating Importance Alignment, BERT-base had the highest importance alignment score. We also found that the smaller, 'base' versions of transformers tended to have higher importance alignment scores than the corresponding large versions. However, being smaller doesn't always result in higher importance alignment, since both small non-transformer models had lower importance alignment scores than 'base' transformers. Finally, we demonstrate that more accurate models (both for classic test accuracy and on the diagnostic dataset HANS; McCoy et al. 2019) 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

The term "alignment" has been used in several different contexts in AI: alignment of model behaviour with normative notions of human ethics ("value alignment"; Russell et al. 2015; Peng et al. 2020), alignment between tokens from source to target in machine translation, alignment between images and text in image-caption alignment models, etc. In this paper, we propose a new type of alignment, Importance Alignment: we want models to not only generate accurate outputs, but also to generate these accurate outputs for reasons that align with human expectations.

High importance alignment can be valuable because prior work has demonstrated that when people form correct mental models of AI decision boundaries, they make better AI-assisted decisions (Bansal et al., 2019a,b). For example, when annotators are provided with additional information about model decisions, such as model accuracy (cf. Yin et al. 2019) or model-generated explanations (Bansal et al., 2020b), it increases their level of trust and in some settings, can actually improve AI-assisted decision making (Zhang et al., 2020). Therefore, optimizing models to have high importance alignment is a worthy goal, even if it can initially result in models with lower accuracy. 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.

3 Measuring Importance Alignment

We assume that for some input example $x = \{x_1, x_2 ... x_n\}$ with n tokens and a gold label y, there exists an annotator-generated explanation of why the gold label is correct and/or why a model might output an incorrect prediction. We convert this natural language explanation into an **oracle importance score** ($\mathbb{I}_m^O(x, y)$) which quantifies the extent to which annotators expect each token in x to push the model's prediction towards or away from the gold label.[1] Then, to compute Importance Alignment we correlate the oracle importance score for x with the **model importance** ($\mathbb{I}_m(x, y)$), which quantifies the extent to which each token in x actually pushes the model's prediction towards or away from the gold label. A greater correlation between model and oracle importance scores indicates a greater alignment between how annotators expect models

[1] We refer to this as an "oracle", because we consider importance scores derived from human-generated explanations to be the ground truth.

to make decisions and how these models actually make decisions. In the remainder of this section we describe our methods to calculate oracle and model importance scores as well as the importance alignment metric from correlating these two scores.

3.1 Computing model importance scores

We compute model importance scores using *Integrated Gradients* (IG; Sundararajan et al. 2017). Concretely, we define model importance ($\mathbb{I}$) for some model m and some example x (e.g., the concatenation of the premise and hypothesis for NLI) with the gold label y, as follows,

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

where IG_m returns a vector of IG attribution scores with respect to the gold label for each token in x and $|\cdot|$ denotes component-wise absolute value.

For some token in the input x_i, a positive IG attribution score indicates that x_i pushed the model's prediction towards the gold label, whereas a negative IG attribution score indicates that x_i pushed the model's prediction away from the gold label. In a model with high importance alignment, we would expect positive attribution scores to be correlated with annotator expectations about why the gold label is correct, and negative attribution scores to be correlated with annotator expectations about why a model might output an incorrect prediction. In this paper, we are considering explanations which capture both of these annotator expectations without differentiating between them. Therefore, we use the absolute value of the IG attribution score.

Why Integrated Gradients? We use Integrated Gradients because they are axiomatically both interpretable and faithful (Sundararajan et al., 2017), unlike attention based methods which have been argued are not faithful explanations of models' decision making processes (Jain and Wallace 2019, but see Wiegreffe and Pinter 2019 for a counterpoint). Other perturbation methods which are more faithful than attention such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), and their variants can be used, although these methods have been argued to be unreliable (Camburu et al., 2019; Slack et al., 2020; Camburu et al., 2021). No current consensus exists on which methods should be employed (Hase and Bansal, 2020; Ghorbani et al., 2019). 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 Computing oracle importance scores

We describe three methods of converting natural language explanations into oracle importance scores: hard oracle, soft oracle and expert oracle.

Hard oracle importance. Hard oracle importance is a binary measure of token overlap between the input and the explanation. This measure captures the intuition that if annotators thought that specific tokens in the input are important for pushing the model towards or away from the gold label, then they will use those words in their explanation. Formally, for some input x with gold label y and explanation e, we define hard oracle importance as follows, where the overlap function yields a binary vector in which the i-th component is valued 1 if $x_i \in e$ and x_i is not a stop word;[2] the i-th component of the vector is valued 0 otherwise.

$$\mathbb{I}^H_m(x, y, e) = overlap(x_m, e) \quad (2)$$

This measure is model specific only to the extent that the models differ in how they tokenize the input x; we assume that the annotator generated explanations themselves do not describe expectations about specific models.

Soft oracle importance. The hard oracle is very simple and does not capture synonyms, entities referred to with pronouns, paraphrases, etc. To overcome these shortcomings, we also define soft oracle importance where we compute importance scores from IG of explanation-informed models. The input for these explanation-informed models is the original input x concatenated with some explanation e. The task of the model is to predict not only the gold label for the original task y, but also a binary output indicating whether e was an explanation about the current example or about some other example. This task requires the model to perform not only the target task (e.g., NLI) but also requires the model to establish a relationship between the provided explanation and the input, thereby incorporating information from the natural language explanation.

[2] We used the list of stop words from NLTK (Bird et al., 2009)

Formally, for some input x with gold label y and explanation e, we define soft oracle importance as,

$$\mathbb{I}_m^S(x, y, e) = |IG_{m'}(x, e, y^{'})| \tag{3}$$

where $m^{'}$ refers to the explanation-informed model and $y^{'}$ refers to the target output of $m^{'}$ which incorporates both y and a binary output indicating whether e is a matched explanation (e.g., for NLI, $y^{'}$ would have six possible values). This measure is also model specific both because of tokenization and because the explanation-informed model has the same model architecture as the target model.

Expert oracle importance. The hard and soft oracles are automatic ways of computing oracle importance scores. To validate these automatic measures, we also computed oracle importance scores from experts (three of the authors on this paper). Given the input, gold label, and annotator generated explanation for a given example, the expert annotators ($N = 3$) indicated which tokens of the input they believed that the original annotator (i.e., the one who generated the explanation) thought were important for the model's prediction. Since generating expert annotations was very time consuming, we computed this measure only for a random subset of 60 examples.

Formally, for some input x with gold label y and explanation e, we define expert oracle importance for any given token as the proportion of expert annotators who indicated that the token was important according to the annotator who generated the explanation. This is expressed as,

$$\mathbb{I}_m^E(x, y, e) = \frac{1}{N}\sum_{k}^{N} expert_k(x_m, e, y) \tag{4}$$

where $expert_k$ returns a binary vector in which the i-th element is valued 1 if annotator k indicated that the i-th token was important, and 0 otherwise.

Like with hard oracle importance, this measure is model specific only to the extent that the models differ in how they tokenize the input x.

We could not compute oracle importance scores from the original annotators and had to rely on expert annotators because this information was absent from the dataset of natural language explanations we used. Additionally, we argue below that collecting high-quality oracle importance annotations from naïve annotators can be very tricky.

Why start from human-generated natural language explanations? We convert natural language explanations to oracle importance scores instead of collecting oracle importance scores directly from naïve annotators for two reasons. First, there already exist data sets of natural language explanations, where annotators were required to reason about models' decision making in an adversarial setting (Nie et al., 2020), and more such data sets are being generated (Kiela et al., 2021). Second, we contend 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?". In fact, to pursue this angle assiduously, one would ideally recruit annotators who know what IG is and ask them to predict IG scores—after all, since we are using IG scores to quantify how the models make decisions, the best way to quantify "how humans think models make decisions" would be to measure what humans think the IG scores will be. Unfortunately, such a task would be challenging for most non-expert annotators, making it infeasible to collect high quality annotations.

3.3 Importance Alignment Metric

For each example with input x, gold label y and explanation e, we compute importance alignment for some model m as the mean Fisher-transformed product-moment (i.e., Pearson's) correlation (r) between the model importance and oracle importance scores for that example as below, where the oracle O is either the hard (H), soft (S) or expert (E) oracles:

$$C_m(x, y) = \operatorname{arctanh}(r(\mathbb{I}_m(x, y), \mathbb{I}_m^O(x, y, e))) \tag{5}$$

We Fisher-transform the correlation coefficient r to ensure that $C_m(x, y)$ is unbiased and does not violate normality assumptions required for the statistical analyses we use (Fisher, 1921).[3]

For each example, we also compute a random baseline for $C_m(x, y)$ where the oracle

[3] Fisher transformed correlation coefficients are approximately normally distributed when the correlation coefficients are calculated from sample pairs drawn from bivariate normal distributions. Although $\mathbb{I}_m(x, y)$ and $\mathbb{I}_m^O(x, y, e))$ are not normally distributed, visual examination of the resulting fisher transformed correlations revealed that $C_m(x, y)$ were in fact approximately normal.

importances are calculated by pairing the input with an explanation (e_R) which was written for a different input example and was chosen at random:

$$C_{m_R}(x, y) = \text{arctanh}(r(\mathbb{I}_m(x, y), \mathbb{I}_m^O(x, y, e_R))) \tag{6}$$

We compute this measure to control for spurious patterns that can drive the correlation between model and oracle importances: we can find a non-zero correlation between the importance scores if a certain model m, such as BERT, (explanation informed or not) always assigned high IG attribution values to tokens at specific indices; we can also find a non-zero correlation if certain tokens (e.g., "the") received high attribution irrespective of the context, and these tokens occurred frequently in the input and explanation.

To measure the extent to which model and oracle importances are correlated with each other over and above spurious correlations, we measure the mean difference ($\Delta\mathcal{A}$) between $C_m(x, y)$ and $C_{m_R}(x, y)$ for all examples in some dataset D and back-transform it to the correlation scale:

$$\Delta\mathcal{A} = \tanh\left(\frac{1}{\mid D \mid}\sum_{(x,y)\in D} C_m(x, y) - C_{m_R}(x, y)\right) \tag{7}$$

To measure whether $\Delta\mathcal{A}$ is significantly greater than 0, we use paired t-test between $C_m(x, y)$ and $C_{m_R}(x, y)$ for all examples in D. We compute a different measure of $\Delta\mathcal{A}$ for each oracle importance score.

4 Experimental details

4.1 Models

Target models. We measured the importance alignment for six pretrained Transformer language models: 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). We fine-tuned these models on the combination of the following NLI datasets: SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), NLI-recast FEVER (Thorne et al., 2019) and ANLI rounds 1–3 (Nie et al., 2020). We used the hyperparameters in the ANLI codebase for fine-tuning.[4]

We used the same datasets to also train two non-transformer models: a bag-of-words (BOW) model and a RNN based model we call **BInferSent** for 'BERT-InferSent'. The BOW model has a single max pooling layer on top of BERT token embeddings. The BInferSent model combines the InferSent architecture of Conneau et al. (2017) with Short-Stacked Sentence Encoders of Nie and Bansal (2017), using 3 layers of BiLSTMs (Hochreiter and Schmidhuber, 1997) with residual connections on top of BERT token embeddings.

	Target models	Explanation informed models
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
BInferSent	40.00	20.97
BOW	35.82	18.70

Table 1: Accuracy on the development partition of the ANLI dataset for target models (finetuned on MNLI + SNLI + ANLI + re-cast FEVER) and models used as the soft oracle (finetuned on 6-way NLI and reason classification on a subset of ANLI).

Explanation informed models. To compute soft oracle importance, we trained explanation informed models for each target model architecture on a six-way classification task. We trained three models per architecture using different random seeds. In this task, the input to the model was a context-hypothesis pair concatenated with an annotator-generated explanation. The output of the model was a joint label indicating whether the context entails, contradicts or is neutral with respect to the hypothesis, and whether the explanation matches the context-hypothesis pair.

We generated the training ($n = 19043$) and development ($n = 2116$) datasets for these classifiers by subsetting the portion of the ANLI training set for which an explanation was provided — i.e., the examples in the training set in which the ANLI annotators had successfully fooled the model. We presented each of the 19043 examples twice when training the explanation informed models: once with a matched explanation (i.e., the original one written by the annotator for that explanation) and once with a randomly selected explanation (see §3.3 above). We trained all models for two epochs.

The accuracy of these explanation informed models either matched or surpassed that of the target model despite being trained on only a

[4] https://github.com/facebookresearch/anli/blob/master/script/example_scripts/

Model	Importance Alignment		Acc.
	$\Delta\mathcal{A}^H$	$\Delta\mathcal{A}^S$	ANLI
BERT-Base	0.21^{***}	0.11^{***}	48.02
RoBERTa-Base	0.11^{***}	0.02^{*}	50.47
ELECTRA-Base	0.17^{***}	0.06^{***}	52.33
BERT-Large	0.18^{***}	-0.02	49.24
RoBERTa-Large	0.04^{*}	0.01	55.37
ELECTRA-Large	0.07	0.01	58.06
All Base Trans.	0.17	0.07	50.27
All Large Trans.	0.11	-0.003	54.56
BInferSent	0.12^{***}	<0.01	40.00
BOW	0.01^{***}	0.01^{***}	35.82

Table 2: 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. $\Delta\mathcal{A}$ **was computed over the examples that the models got wrong**. Average model accuracy across seeds and different rounds of ANLI is also provided. '*'s indicate whether $\Delta\mathcal{A}$ is significantly greater than 0. '***' indicates $p < 0.001$, '**' indicates $p < 0.01$ and '*' indicates $p < 0.05$.

Model	Importance Alignment		
	$\Delta\mathcal{A}^H$	$\Delta\mathcal{A}^S$	$\Delta\mathcal{A}^E$
BERT-Base	0.21^{***}	0.14^{***}	0.39^{***}
RoBERTa-Base	0.10^{***}	0.04^{***}	0.31^{***}
ELECTRA-Base	0.15^{***}	0.09^{***}	0.33^{***}
BERT-Large	0.15^{***}	-0.03	0.27^{***}
RoBERTa-Large	0.02	-0.01	0.21^{***}
ELECTRA-Large	0.08	0.02	0.20^{*}
All Base Trans.	0.16	0.09	0.34
All Large Trans.	0.09	-0.01	0.23
BInferSent	0.14^{***}	-0.003^{***}	0.29^{***}
BOW	0.02	0.03	0.04

Table 3: Importance Alignment between model importance scores and oracle importance scores ($\mathcal{A}^H$, $\mathcal{A}^S$ and $\mathcal{A}^E$ metrics) across 5 random seeds on the ANLI dataset. $\Delta\mathcal{A}$ **was computed over the 60 examples with expert annotations**. '*'s indicate whether $\Delta\mathcal{A}$ is significantly greater than 0. '***' indicates $p < 0.001$, '**' indicates $p < 0.01$ and '*' indicates $p < 0.05$.

small subset of the original data (see Table 1). This suggests that these models did learn to incorporate information about explanations, and that the information present in the explanations was useful for NLI.

4.2 Evaluation Datasets

ANLI. We measured importance alignment on the development set of the ANLI dataset. In this dataset, 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 the provided label was correct and why they thought model was fooled into generating an incorrect prediction. While other datasets with natural language explanations for NLI exist, (e.g., e-SNLI; Camburu et al. 2018), the explanations in these datasets only address why the gold label is correct, and not why the annotators thought the model generated an incorrect prediction. Additionally, the adversarial setting in which ANLI was collected encourages the annotators to reason about models' decision making. These two factors make ANLI more suitable for our purposes than other explanation datasets.

Since annotators provided explanations only when the model generated an incorrect prediction, to allow for an apples-to-apples comparison, we only compute importance alignment for examples in which our target models generate an incorrect prediction. As a consequence, the specific examples used to measure importance alignment differs across models: different models fail on different examples. To ensure that our results are not driven by the differences between examples, we repeated our analyses on a subset of examples that all models failed on, and we found that our results were nearly identical (compare Table 2 and Table 3).

We collected expert annotations (see §3.1) on 60 randomly sampled examples from the set of examples that all the models failed on, 20 from each round of ANLI. We computed the expert importance alignment score ($\mathcal{A}^E$) for this subset of examples and repeated our analyses.

HANS. We measured the extent to which importance alignment scores correlated with accuracy on the HANS diagnostic dataset (McCoy et al., 2019). This dataset measures the extent to which NLI models rely on non-human like heuristics when performing inference, such as inferring that the context entails the hypothesis if all of the words in the hypotheses are in the context: models which do not rely on these non-human like heuristics have higher accuracy on the this dataset. If we assume that naïve annotators in general do not expect NLI models to rely on such heuristics, then we might expect models with high importance alignment (i.e., models which make inference decisions in ways annotators expect them to) to also have high accuracy on the HANS dataset.

5 Results

In Table 2, we report the importance alignment scores for the hard and soft oracles computed over all the examples the models generated an incorrect prediction for, averaged across the five random seeds. In Table 3, we report the importance alignment scores for hard, soft and expert oracles computed over the subset of 60 examples with expert annotations. Since the importance alignment scores for the hard and soft oracles are nearly identical across both tables, we focus our discussion of the results from the subset of examples with expert annotations.We repeated the reported analyses with hard and soft oracles on all the examples, and found qualitatively similar results (see Appendix A).

Effect of model size on Importance Alignment. Across all three types of Importance Alignment scores ($\mathcal{A}^H$, $\mathcal{A}^S$ and $\mathcal{A}^E$) the base versions of the transformer models had higher importance alignment than their larger counterparts, with BERT-base having the highest importance alignment. To test the statistical significance of this this observation, we fit three linear mixed effect regression models: one for each type of oracle.

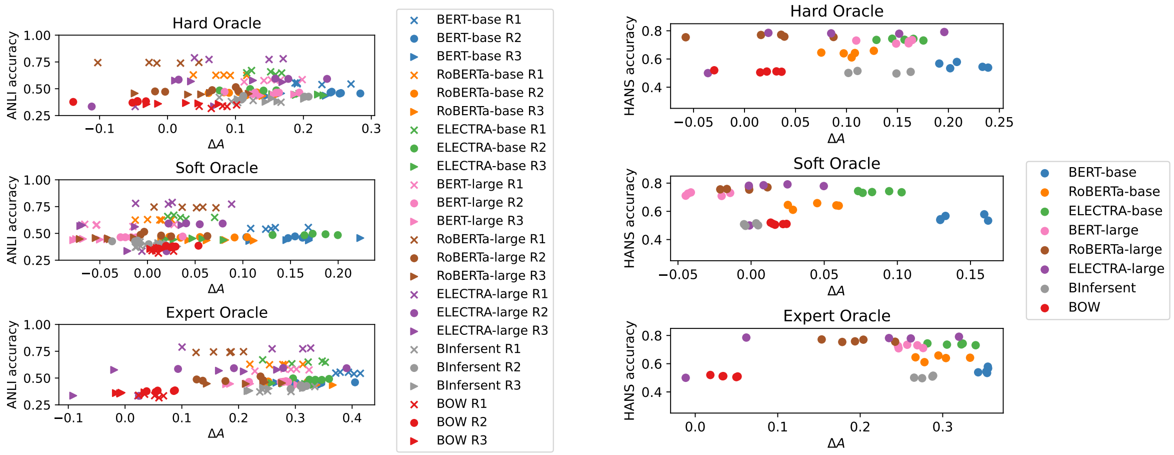
We predicted the pair-wise difference between $C_m(x, y)$ and $C_{m_R}(x, y)$ (see Equation 5 and Equation 6) as a function of the following predictors: model size (base vs. large), model type (BERT vs. ELECTRA and BERT vs. RoBERTa) and the interaction between the two. We included model type and its interaction with model size as predictors to measure the effect of model size over and above the differences between specific models. We also included a random intercept and random slope of model size for every example to incorporate the following assumptions: first, the difference $C_m(x, y)$ and $C_{m_R}(x, y)$ can differ for every example; second, the difference between base and large models can also differ for every example. The results described below were significant at a threshold of $p < 0.005$ unless specified otherwise. For further details see Appendix A.

The analyses indicated that across all measures importance alignment, $\Delta\mathcal{A}$ was significantly greater for the base models when compared to their larger counterparts. Additionally, $\Delta\mathcal{A}$ was significantly greater in BERT models than in RoBERTa and ELECTRA models (base and large).

Although the results suggest that smaller models have stronger importance alignment than their larger counterparts, our experiments with the BInfersent and BOW models suggest that smaller models do not always have higher importance alignment: the importance alignment for both these models is lower than the alignment for BERT-base model (the smallest of the transformer models when taking into consideration both the number of parameters and pre-training size).

To test the statistical significance of this numerical result, we fit another set of linear mixed effects model where we predicted the pair-wise difference between $C_m(x, y)$ and $C_{m_R}(x, y)$ as a function of model type (BERT-base vs. BInferSent and BERT-base vs. BOW). We also included a random intercept of item. As expected, the importance alignment for the BERT-base model was significantly greater than that for the BInferSent and BOW models for all the measures.

ANLI accuracy and Importance Alignment. Based on the results from Table 2 and Table 3, we wondered whether importance alignment might be negatively correlated with NLI accuracy: the base versions of the models which have higher importance alignment have lower accuracy on ANLI compared to their larger counterparts. To test this, we computed a separate value of $\mathcal{A}^H$, $\mathcal{A}^S$ and $\mathcal{A}^E$ for each random seed of the target model and for each round of ANLI. Then, we computed the product moment (i.e., Pearson's) correlation between model accuracy on NLI and $\mathcal{A}^H$, $\mathcal{A}^S$ and $\mathcal{A}^E$. We found a significant correlation between accuracy and $\mathcal{A}^E$ ($r = 0.42$; $p = 0.008$). However, this correlation was being driven entirely by the BOW models, which had both low accuracy and low importance alignment scores. When repeating the analysis with the BOW models excluded, we found that none of the measures of importance alignment were significantly correlated with NLI accuracy ($\mathcal{A}^H$: $r = -0.09$ and $p = 0.61$; $\mathcal{A}^S$: $r = 0.04$ and $p = 0.80$; $\mathcal{A}^E$ $r = -0.04$ and $p = 0.83$; see Figure 2a). We repeated the analyses for $\mathcal{A}^H$ and $\mathcal{A}^S$ for all wrong examples and found that only $\mathcal{A}^H$ was significantly correlated with accuracy ($r = 0.39$, $p = 0.01$; see Appendix A.2)



(a) Accuracy on ANLI is not correlated with importance alignment ($\mathcal{A}^H$ and $\mathcal{A}^S$). The cross, circle, and triangle refer to rounds 1, 2, and 3 of ANLI, respectively.

(b) Accuracy on HANS is not correlated with importance alignment. Alignment values are averaged across rounds because HANS is not divided into rounds.

HANS accuracy and Importance Alignment. As discussed earlier, we hypothesized that high importance alignment might result in models relying less on non-human-like heuristics, thereby resulting in higher accuracy on the HANS dataset. We found no such correlation, however ($\mathcal{A}^H$: $r = 0.04$ and $p = 0.82$; $\mathcal{A}^S$: $r = -0.21$ and $p = 0.21$; $\mathcal{A}^E$ $r = 0.23$ and $p = 0.17$; see Figure 2b).[5]

This lack of correlation is likely a result of the mismatch between how human-likeness is defined in HANS and in our importance alignment measures. In HANS, the targeted heuristics are simple (i.e., can be articulated with a rule), and describe general principles of how models ought not behave if they are to be human-like. In contrast, our measures of importance alignment are derived from example level explanations of how naïve annotators expected models to behave, and as such are not based on any overarching easy-to-articulate principles. When evaluating whether models make decisions as humans expect them to, jointly considering both these definitions of human-likeness can be useful.

Comparing the Importance Alignment Scores. We used hard and soft oracles to automatically measure oracle importance scores. To validate these methods, we computed the Spearman rank correlation between the importance scores derived from these methods and from the manually annotated expert oracle. The hard oracle was moderately correlated with the expert oracle ($r = 0.24$, $p < 0.0001$), whereas the soft oracle was more weakly correlated ($r = 0.14$, $p < 0.0001$). Additionally, the hard and soft oracle importance scores were also weakly correlated with each other ($r = 0.11$, $p < 0.0001$).[6] Taken together, these results suggest that neither the hard nor the soft oracle measures are perfect proxies for expert human importance scores. This imperfection does not impact the conclusions we draw in this paper, however: the results we discussed held true across all the measures.

6 Discussion

In this paper, we argued that it is important to not only evaluate models on how accurate they are on a given task, but also on whether the decisions that the models make align with how humans expect them to make these decisions. We introduced a measure called Importance Alignment which quantifies the extent to which the parts of the input that non-expert annotators expected to influence models' decisions actually influenced the decisions. To quantify which parts of the input influenced model decisions (*model importance*), we used Integrated Gradients. To quantify annotator expectations (*oracle importance*), we proposed three methods of quantifying annotator generated natural language explanations of model behaviour: two automatic and one that relied on expert input.

As a case study, we applied this method to measure Importance Alignment in eight NLI models (six transformers, an LSTM and a BOW model), using annotator generated explanations from the ANLI dataset. We found that, across

[5] When we considered the heuristics separately, we found some marginally significant correlations (see Appendix A.2)

[6] The results are comparable with Pearson's correlation.

all three measures of importance alignment, the base versions of the transformer models had significantly higher importance alignment than their larger counterparts, with BERT-base having the highest importance alignment. Smaller models do not always result in higher importance alignment, however: the BERT-base model had higher alignment than the LSTM and BOW models. Additionally, importance alignment scores were not correlated with model accuracy on ANLI or the HANS diagnostic dataset in most cases. This suggests that importance alignment and accuracy are complementary methods of evaluating models.

Future work. There are at least four directions in which this work can be extended. First, future work can evaluate whether our conclusions about model size in NLI models generalizes to smaller transformer models (Turc et al., 2019; Warstadt et al., 2020) and to models trained on other NLP tasks.

Second, future work can build on our methods of calculating model and oracle importance. The two methods of automatically computing oracle importance we proposed as a starting point were only moderately correlated with the method that relies of expert input. Future work can develop better methods of measuring oracle importance incorporating the strengths of both. Future work can also explore other existing ways of calculating model importance (e.g., LIME, SHAP and their variants).

Third, future work can generate more detailed datasets of natural language explanations of model behaviour. For example, in the dataset we used, the explanations the annotators provided were an amalgamation of both why the the label was correct and why the model might have been fooled. Additionally, annotators generated explanations only when the model generated an incorrect output. By disentangling these two types of explanations and collecting explanations for when models generate correct outputs future work can separately study whether models succeed and fail in ways people expect them to. Similarly in the dataset we used, each context-hypothesis-label triplet was associated with only one annotator generated explanation. However, it is possible that there are several explanations of model behaviour that are equally valid. Collecting more explanations per triplet can improve our understanding of how people expect models to succeed and fail.

Fourth, future work could explore which factors drive higher importance alignment. For example, we observed that transformer models with fewer parameters had higher importance alignment than models with more parameters. Does this finding apply to non-transformer architectures too? Is there a threshold for model parameters, where this inverse relationship between model size and alignment score breaks down? Additionally, can other factors like model architecture, type of training objective or the types of sentences the models were trained on influence importance alignment? Such exploration can not only result in models better aligned with human expectations, but also improve our understanding of what drives importance alignment.

7 Conclusion

We proposed a novel metric, Importance Alignment, to measure the extent to which human-generated explanations of model decisions align with how models actually make these decisions. As a case study, we used this metric to evaluate eight different models trained on NLI and found that the BERT-base model had the highest alignment with human-generated explanations. We also found that our metric was not correlated with model accuracy, suggesting that accuracy and Importance Alignment are complementary ways of evaluating models.

Acknowledgements

We would like to thank Luke Zettlemoyer, Ana Valeria Gonzalez, the Dynabench team, Roy Schwartz, Oana Maria Camburu, Tom McCoy, and Tal Linzen's Computation and Psycholinguistics lab for their invaluable comments and support.

References

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. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):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.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

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, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. 2019. Can I trust the explainer? verifying post-hoc explanatory methods. In *NeurIPS 2019 Workshop on Safety and Robustness in Decision Making*, Vancouver, Canada.

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. 2021. The struggles of feature-based explanations: Shapley values vs. minimal sufficient subsets. In *Proceedings of the Explainable Agency in Artificial Intelligence Workshop at AAAI 2021*.

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.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

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. *arXiv*.

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 and Mohit Bansal. 2017. Shortcut-stacked sentence encoders for multi-domain inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 41–45, Copenhagen, Denmark. 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.

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).

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962.

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.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning helpful inductive biases from self-supervised pretraining. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

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

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.

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*.

A Details about statistical analyses

A.1 Effect of model size

Formula for Transformer models

```
smf.mixedlm("fisher_cor_diff ~ C(model_size, Treatment(reference=’base’))
                              *C(model,Treatment(reference=’bert’))",
            data = transformer_dat, groups = transformer_dat["ex_id"],
            re_formula="~model_size")
```

Formula for smaller models (BERT-base, InferSent and BOW)

```
smf.mixedlm("fisher_cor_diff ~ C(model,Treatment(reference=’bert’))",
             data = small_dat, groups = small_dat["ex_id"],
             re_formula="~model_size")
```

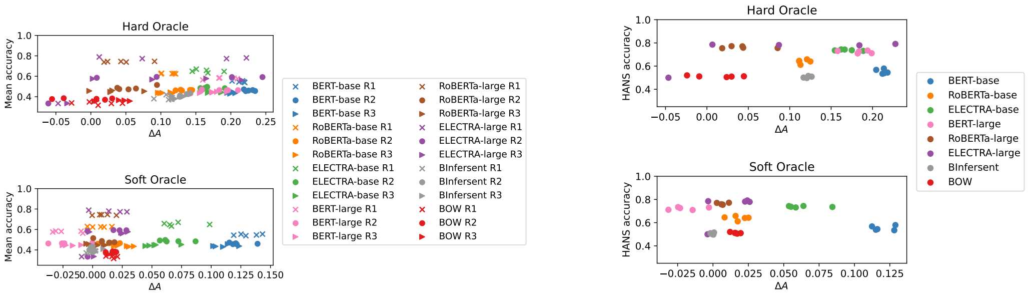
Fit	Coefficient	All wrong examples		Expert annotated subset		
		$\Delta\mathcal{A}^H$	$\Delta\mathcal{A}^S$	$\Delta\mathcal{A}^H$	$\Delta\mathcal{A}^S$	$\Delta\mathcal{A}^E$
Transformer models	Large (vs. Base)	-0.32***	-0.142***	-0.057**	-0.173***	-0.115***
	ELECTRA (vs. BERT)	-0.044***	-0.057***	-0.048 **	-0.067***	-0.054*
	RoBERTa (vs. BERT)	-0.099***	-0.101***	-0.106***	-0.105***	-0.083***
	ELECTRA : Large	-0.070***	0.085***	-0.014	0.109***	-0.051
	RoBERTa : Large	-0.044***	0.127***	-0.016	0.134***	0.012
Smaller models	BOW (vs. BERT-base)	-0.203***	-0.104***	-0.184***	-0.125***	-0.184***
	BInferSent (vs. BERT-base)	-0.094***	-0.120***	-0.080***	-0.154***	-0.080***

Table 4: ***, **and *indicate p < 0.0001, 0.001 and 0.01 respectively and ‘:’ indicates an interaction effect. A separate mixed effects regression model was fit for each column. Negative coefficients for the main effects indicate that the baseline value was greater than the comparison (e.g., Base was greater than Large).

A.2 Correlation with accuracy

Evaluation dataset	All wrong examples		Expert annotated subset		
	$\Delta\mathcal{A}^H$	$\Delta\mathcal{A}^S$	$\Delta\mathcal{A}^H$	$\Delta\mathcal{A}^S$	$\Delta\mathcal{A}^E$
ANLI Dev	0.39^*	0.10	0.23	0.10	0.42^{**}(-0.04)
HANS (all)	0.18	-0.19	0.04	-0.21	0.22
HANS (constituent)	0.30^+	-0.11	-0.19	-0.26	-0.18
HANS (lexical overlap)	0.20	-0.17	-0.25	-0.27	-0.19
HANS (subsequence)	-0.006	-0.31^+	-0.29^+	-0.39^*	-0.27

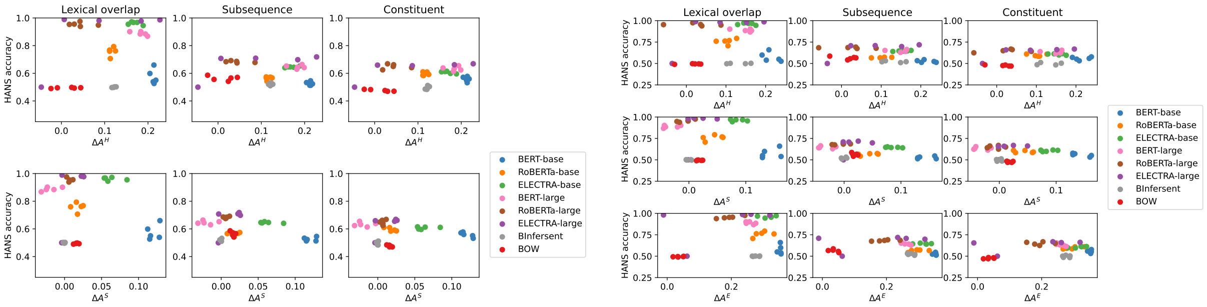
Table 5: '+', '*' and '**' indicate p < 0.1, 0.05 and 0.01 respectively. Values in parentheses indicate correlation without the BOW models in cases where they were outliers (see Figure 2a)



(a) Correlationg accuracy on ANLI with importance alignment ($\mathcal{A}^H$ and $\mathcal{A}^S$) for all wrong examples. The cross, circle, and triangle refer to rounds 1, 2, and 3 of ANLI, respectively.

(b) Correlating accuracy on HANS with importance alignment for all wrong examples. Alignment values are averaged across rounds because HANS is not divided into rounds.

Figure 3



(a) Correlating accuracy on ANLI with importance alignment ($\mathcal{A}^H$ and $\mathcal{A}^S$) for all wrong examples. The cross, circle, and triangle refer to rounds 1, 2, and 3 of ANLI, respectively.

(b) Correlating accuracy on HANS with importance alignment for all wrong examples. Alignment values are averaged across rounds because HANS is not divided into rounds.

Figure 4