Connecting Attributions and QA Model Behavior on Realistic Counterfactuals

Xi Ye **Rohan Nair** **Greg Durrett**
Department of Computer Science
The University of Texas at Austin
{xiye,rnair,gdurrett}@cs.utexas.edu

Abstract

When a model attribution technique highlights a particular part of the input, a user might understand this highlight as making a statement about counterfactuals (Miller, 2019): if that part of the input were to change, the model's prediction might change as well. This paper investigates how well different attribution techniques align with this assumption on *realistic* counterfactuals in the case of reading comprehension (RC). RC is a particularly challenging test case, as token-level attributions that have been extensively studied in other NLP tasks such as sentiment analysis are less suitable to represent the reasoning that RC models perform. We construct counterfactual sets for three different RC settings, and through heuristics that can connect attribution methods' outputs to high-level model behavior, we can evaluate how useful different attribution methods and even different formats are for understanding counterfactuals. We find that pairwise attributions are better suited to RC than token-level attributions across these different RC settings, with our best performance coming from a modification that we propose to an existing pairwise attribution method.[1]

1 Introduction

Recent research in interpretability of neural models (Lipton, 2018) has yielded numerous post-hoc explanation methods, including token attribution techniques (Ribeiro et al., 2016; Sundararajan et al., 2017; Guan et al., 2019; De Cao et al., 2020). Attributions are a flexible explanation format and can be applied to many domains, including sentiment analysis (Guan et al., 2019; De Cao et al., 2020), visual recognition (Simonyan et al., 2013), and natural language inference (Camburu et al., 2018; Thorne et al., 2019). However, it is hard to evaluate whether these explanations are *faithful* to the computation of the original model (Wu and Mooney, 2019; Hase and Bansal, 2020; Wiegreffe et al., 2021; Jacovi and Goldberg, 2020) and as a result, they can potentially mislead users (Rudin, 2019). Furthermore, attributions do not have a consistent and meaningful *social attribution* (Miller, 2019; Jacovi and Goldberg, 2021): that is, when a user of the system looks at an explanation, they do not necessarily draw a valid conclusion from it, making it hard to use for downstream tasks.

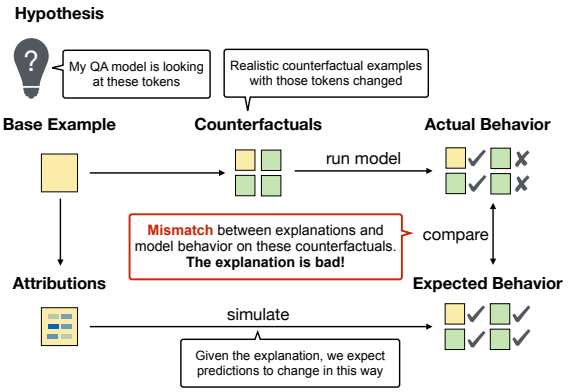


Figure 1: Our methodology. Given a base example, we can formulate a hypothesis about the model's behavior, like a theory about how the model is using certain tokens. Next, we collect counterfactual examples that modify these tokens and profile the actual model behavior. Finally, we assess whether feature attributions suggest behavior consistent with what we observe, verifying whether our attributions actually enable meaningful statements about behavior on counterfactuals.

How can we evaluate whether these attributions make faithful and meaningful statements about model behavior? In this work, we show how to use counterfactual examples to evaluate attributions' ability to reveal the **high-level** behavior of models. That is, rather than a vague statement like "this word was important," we want attributions to give concrete, testable conclusions like "the model compared these two words to reach its decision;" this statement can be evaluated for faithfulness and

[1] Code and data: https://github.com/xiye17/EvalQAExpl

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5496–5512
November 7–11, 2021.

it helps a user make important inferences about how the system behaves. We approach this evaluation from a perspective of simulatability (Hase and Bansal, 2020): can we predict how the system will behave on new or modified examples? Doing so is particularly challenging for the RC models we focus on in this work due to the complex nature of the task, which fundamentally involves a correspondence between a question and a supporting text context.

Figure 1 shows our methodology. Our approach requires annotating small sets of realistic counterfactuals, which are perturbations of original data points. These resemble several prior "stress tests" used to evaluate models, including counterfactual sets (Kaushik et al., 2020), contrast sets (Gardner et al., 2020), and checklists (Ribeiro et al., 2020). We first semi-automatically curate these sets to answer questions like: if different facts were shown in the context, how would the model behave? If different amounts of text or other incorrect paragraphs were retrieved by an upstream retrieval system, would the model still get the right answer?

We run the model on counterfactuals to assess the ground truth behavior. Then, given attributions from various techniques, can we predict how the model would behave **based purely on these explanations**? Our approach to do this is specific to each dataset and attribution method, but generally involves assessing how strongly the attribution method highlights tokens that are counterfactually altered, which would indicate that those tokens should impact the prediction if changed.

To showcase the kind of evaluation this method can enable, we investigate two paradigms of explanation techniques: token attribution-based (Simonyan et al., 2013; Ribeiro et al., 2016; De Cao et al., 2020) and feature interaction-based (Tsang et al., 2020; Hao et al., 2021), which attribute decisions to sets of tokens or pairwise token interactions. For both techniques, we devise methods to connect these explanations to our high-level hypotheses about behavior on counterfactual examples. On two types of questions from HotpotQA (Yang et al., 2018) and questions from adversarial SQuAD (Rajpurkar et al., 2016), we show that token-level attribution is not sufficient for analyzing RC models, which naturally involves more complex reasoning over multiple clues. We further propose a modification to an existing interaction technique from Hao et al. (2021) and show improved performance on our datasets.

Our main contributions are: (1) We propose a new goal for attributions, namely automatically simulating model behavior on realistic counterfactuals. (2) We describe a technique for connecting low-level attributions (token-level or higher-order) with high-level model hypotheses. (3) We improve an attention-based pairwise attribution technique with a simple but effective fix, leading to strong empirical results. (4) We analyze a set of QA tasks and show that our approach can derive meaningful conclusions about counterfactuals on each. Overall, we establish a methodology for analyzing explanations that we believe can be adapted to studying attribution methods on a wide range of other tasks with appropriate counterfactuals.

2 Motivation

We start with an example of how model attributions can be used to understand model behavior and consequently how to use our methodology to compare different attribution techniques. Figure 2 shows an example of a multi-hop yes/no question from HotpotQA. The QA model correctly answers *yes* in this case. Given the original example, the explanations produced using IntGrad (Sundararajan et al., 2017) and DiffMask (De Cao et al., 2020) (explained in Section 4) both assign high attribution scores to the two *documentary* tokens appearing in the context: a user of the system is likely to impute that the model is comparing these two values, as it's natural to assume this model is using the highlighted information correctly. By contrast, the pairwise attribution approach we propose in this work (Section 4.3) attributes the prediction to interactions with the question, suggesting the interactions related to *documentary* do not matter.

We manually curate a set of contrastive examples to test this hypothesis. If the model truly recognizes that both movies are documentaries, then replacing either or both of the *documentary* tokens with *romance* should change the prediction. To verify that, we perturb the original example to obtain another three examples (left side of Figure 2). These four examples together form a local neighborhood (Ribeiro et al., 2016; Kaushik et al., 2020; Gardner et al., 2020) consisting of realistic counterfactuals.[2]

[2]One could argue that these counterfactuals are not entirely realistic: a romance film about smoking is unlikely. Generating suitable counterfactuals is a very hard problem (Qin et al., 2019), requiring deep world knowledge of what scenarios make sense or what properties hold for certain entities. The

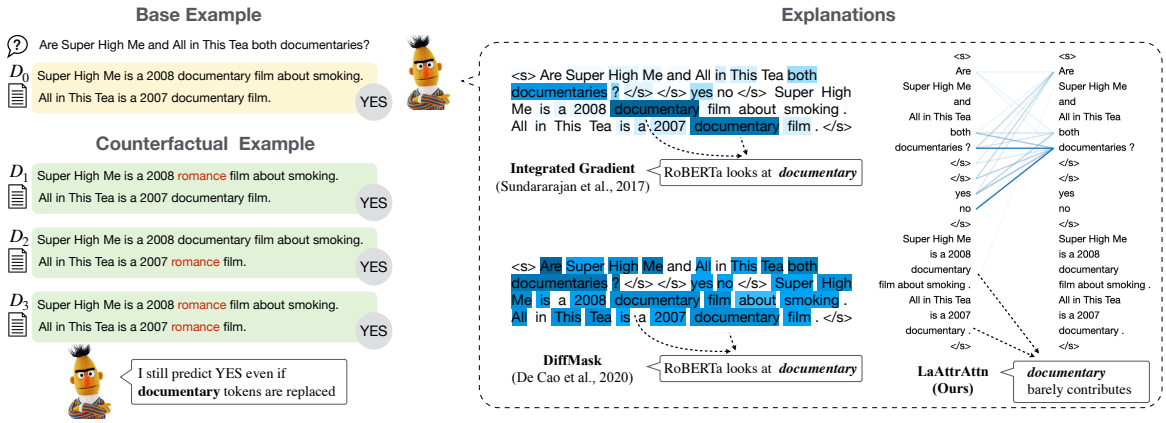


Figure 2: A motivating example and attributions generated by three methods. We profile the model's behavior with its predictions on realistic counterfactual inputs, which suggest the model does not truly base its prediction on the two movies being documentaries. We can evaluate attributions by heuristically assessing whether the attribution mass yields the same conclusion about model behavior.

Unlike what's suggested by the token attribution based techniques, the model always predicts "yes" for every example in the neighborhood, casting doubt on whether the model is following the right reasoning process. Although the pairwise attribution seemed at first glance much less *plausible* than that generated by the other techniques, it was actually better from the perspective of *faithfully* simulating the model's behavior on these examples.

Our main assumption in this work can be stated as follows: **an explanation should describe model behavior with respect to realistic counterfactuals.** Past work has evaluated along plausibility criteria (Lei et al., 2016; Strout et al., 2019; Thorne et al., 2019), but as we see from this example, faithful explanations (Subramanian et al., 2020; Jacovi and Goldberg, 2020, 2021) are better aligned with our goal of simulatability. We argue that a good explanation is one that aligns with the model's high-level behavior, from which we can understand how the model generalizes to new data. How to interpret behavior from explanations is still an open question, but we take initial steps in this work with techniques based on assessing the attribution "mass" on perturbed tokens.

Discussion: Realistic Counterfactuals Many counterfactual modifications are possible: past work has looked at injecting non-meaningful triggers (Wallace et al., 2019), deleting chunks of content (Ribeiro et al., 2016), or evaluating interpolated input points as in IntGrad, all of which violate assumptions about the input distribution. In RC, masking part of the question often makes it nonsensical and we may not have strong expectations about our model's behavior in this case.[3] Focusing on realistic counterfactuals, by contrast, illuminates fundamental problems with our RC models' reasoning capabilities (Jia and Liang, 2017; Chen and Durrett, 2019; Min et al., 2019; Jiang and Bansal, 2019). This is the same motivation as that behind contrast sets (Gardner et al., 2020), but our work focuses on benchmarking explanations, not models themselves.

3 Behavior on Counterfactuals

We seek to formalize the reasoning we undertook in Figure 2. Using the model's explanation on a base data point, can we predict the model's behavior on the perturbed instances of that point?

Definitions Given an original example D_0 (e.g., the top example in Figure 2), we construct a set of perturbations $\{D_1, ..., D_k\}$ (e.g., the three counterfactual examples in Figure 2), which together with D_0 form a local neighborhood $\mathcal{D}$. These perturbations are realistic inputs derived from existing datasets or which we construct.

We formulate a hypothesis $\mathcal{H}$ about the neighborhood. In Figure 2, $\mathcal{H}$ is "the model is comparing the target properties" (*documentary* in this case). Based on the model's behavior on the set $\mathcal{D}$, we can derive a high-level behavioral label z corre-

"true" set of realistic counterfactuals is highly domain-specific, but nevertheless, a good explanation technique should work well on a range of counterfactuals like those considered here.

[3]The exception is in adversarial settings; however, many adversarial attacks do not draw on real-world threat models (Athalye et al., 2018), so we consider these less important.

sponding to the truth of $\mathcal{H}$. We form our local neighborhood to check the answer empirically and compute a ground truth for z. Since the model always predicts "yes" in this neighborhood, we label set $\mathcal{D}$ with $z = 0$ (the model is not comparing the properties). We label $\mathcal{D}$ as $z = 1$, when the model does predict "no" for some perturbations.

Procedure Our approach is as follows:

1. Formulate a hypothesis $\mathcal{H}$ about the model
2. Collect realistic counterfactuals $\mathcal{D}$ to test $\mathcal{H}$ empirically for some base examples
3. Use the explanation of each base example to predict z. That is, learn the mapping $D_0 \rightarrow z$ based on the explanation of D_0 so we can **simulate the model** on $\mathcal{D}$ without observing the perturbations.

Note that this third step *only* uses the explanation of the *base* data point: explanations should let us make conclusions about new counterfactuals without having to do inference on them.

Simulation from attributions In our experiments on HOTPOTQA and SQUAD, we compute a scalar factor f for each attribution representing the importance of a specific part of the inputs (e.g., the *documentary* tokens in Figure 2), which we believe should correlate with model predictions on the counterfactuals. If an attribution assigns higher importance to this information, it suggests that the model will actually change its behavior on these new examples.

Given this factor, we construct a simple classifier where we predict $z = 1$ if the factor f is above a threshold. We expect the factors extracted using better attribution methods should better indicate the model behavior. Hence, we evaluate the explanation using the **best simulation accuracy it can achieve** and the AUC score (S-ACC and S-AUC).[4]

Our evaluation resembles the human evaluation in Hase and Bansal (2020), which asks human raters to predict a model's decision given an example together with its explanations, addressing simulatability from a user-focused angle. Our method differs in that (1) we automatically extract a factor to predict model behavior instead of asking humans to do so, and (2) we predict the behavior on unseen counterfactuals given the explanation of a single base data point.

[4]We do not collect large enough datasets to train a simulation model, but given larger collections of counterfactuals, this is another approach one could take.

4 Explanation Techniques

Compared to classification tasks like sentiment analysis, RC more fundamentally involves interaction between input features, especially between a question and a context. This work will directly compare feature interaction explanations with token attribution techniques that are more common for other tasks.[5]

For RC, each instance $D = (q, c, a)$, a tuple containing a question, context, and answer respectively. In the techniques we consider, q and c are concatenated and fed into a pre-trained transformer model, so our attribution techniques will explain predictions using both of these.

4.1 Token Attribution-Based

These techniques all return scores s_i for each token i in both the question and context.

LIME (Ribeiro et al., 2016) and **SHAP** (Lundberg and Lee, 2017) both compute the attribution values for individual input features by using a linear model to locally approximate the model's predictions on a set of perturbed instances around the base data point. The attribution value for an individual input feature is the corresponding weight of the linear model. LIME and SHAP are different in the way of specifying instance weights used to train the linear model: LIME computes the weights heuristically, whereas SHAP uses a procedure based on Shapley values.

Integrated Gradient (INTGRAD) (Sundararajan et al., 2017) computes an attribution for each token by integrating the gradients of the prediction with respect to the token embeddings over the path from a baseline input (typically mask or pad tokens) towards the designated input. Although a common technique, recent work has raised concern about the effectiveness of INTGRAD methods for NLP tasks, as interpolated word embeddings do not correspond to real input values (Harbecke and Alt, 2020; Sanyal and Ren, 2021).

Differentiable Mask (DIFFMASK) (De Cao et al., 2020) learns to mask out a subsets of the input tokens for a given example while maintaining a distribution over answers as close to the original distribution as possible. This mask is learned in

[5]A potentially even more powerful format would be a program approximating the model's behavior, as has been explored in the context of reinforcement learning (Verma et al., 2018; Bastani et al., 2018). However, beyond limited versions of this (Ribeiro et al., 2018), prior work does not show how to effectively build this type of explanation for QA at this time.

a differentiable fashion, then a a shallow neural model (a linear layer) is trained to recognize which tokens to discard.

4.2 Feature Interaction-Based

These techniques all return scores s_i for each pair of tokens (i, j) in both the question and context that are fed into the QA system.

Archipelago (Tsang et al., 2020) measures non-additive feature interaction. Similar to DIFFMASK, ARCHIP is also implicitly based on unrealistic counterfactuals which remove tokens. Given a subset of tokens, ARCHIP defines the contribution of the interaction by the prediction obtained from masking out all the other tokens, only leaving a very small fraction of the input. Applying this definition to a complex task like QA can result in a nonsensical input.

Attention Attribution (ATATTR) (Hao et al., 2021) uses attention specifically to derive pairwise explanations. However, it avoids the pitfalls of directly inspecting attention (Serrano and Smith, 2019; Wiegreffe and Pinter, 2019) by running an integrated gradients-like procedure over all the attention links within transformers, yielding attribution scores for each link. The attribution scores directly reflect the attribution of the particular attention links, making this model able to describe pairwise interactions.

Concretely, define the h-head attention matrix over input D with n tokens as $A = [A_1, ..., A_l]$, where $A_i \in \mathbb{R}^{h \times n \times n}$ is the attention scores for each layer. We can obtain the attribution score for each entry in the attention matrix A as:

$$\text{ATTR}(A) = A \odot \int_{\alpha=0}^{1} \frac{\partial F(D, \alpha A)}{\partial A} d\alpha, \quad (1)$$

where $F(D, \alpha A)$ is the transformer model that takes as input the tokens and a matrix specifying the attention scores for each layer. We later sum up the attention attributions across all heads and layers to obtain the pairwise interaction between token (i, j), i.e., $s_{ij} = \sum_m \sum_n \text{ATTR(A)}_{\text{mnij}}$.

4.3 Layer-wise Attention Attribution

We propose a new technique LATATTR to improve upon ATATTR for the RC setting. The ATATTR approach simultaneously increases all attention scores when computing the attribution, which could be problematic. Since the attention scores of higher layers are determined by the attention scores of lower layers, forcibly setting all the attention scores and computing gradients at the same time may distort the gradients for the lower level links and produce inaccurate attribution. When applying INTGRAD approach in other contexts, we typically assume the independence of input features (e.g., pixels of an image and tokens of an utterance), an assumption that does not hold here.

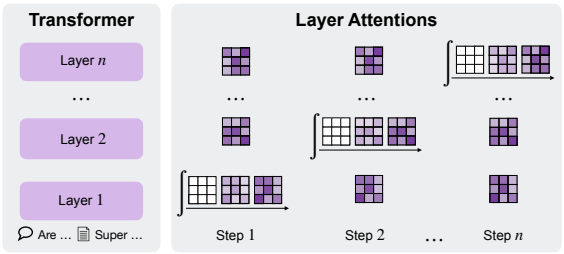


Figure 3: Steps of our Layer-wise Attention Attribution approach, where we modify a single layer's attention at a time. For example, to compute the attribution of attentions at layer 2, we only intervene on the attention matrix at that layer, and leave the other attentions computed as usual.

To address this issue, we propose a simple fix, namely applying the INTGRAD method layer-by-layer. As shown in Figure 3, to compute the attribution for attention links of layer i, we only change the attention scores at layer i:

$$\text{ATTR}(A_i) = A_i \odot \int_{\alpha=0}^{1} \frac{\partial F_{/i}(D, \alpha A_i)}{\partial A_i} d\alpha. \quad (2)$$

$F_{/i}(D, \alpha A_i)$ denotes that we only intervene on the attention masks at layer i while leaving other attention masks computed naturally via the model. We pool to obtain the final attribution for pairwise interaction as $s_{ij} = \sum_m \sum_n \text{ATTR(A)}_{\text{mnij}}$.

5 Experiments

We assess whether attributions can achieve our proposed goal following the setup in Section 3 on the HOTPOTQA dataset (Yang et al., 2018), and the SQUAD dataset (Rajpurkar et al., 2016), specifically leveraging examples from adversarial SQUAD (Jia and Liang, 2017).

Implementation Details For experiments on HOTPOTQA, we base our analysis on a ROBERTA (Liu et al., 2019) QA model in the distractor setting. We implement our model using the Huggingface library (Wolf et al., 2020) and train the model for 4 epochs with a learning rate of 3e-5, a batch size of 32, and a warm-up ratio of 6%. Our model achieves

(a)
Question: Were Ulrich Walter and Léopold Eyharts both from Germany?
Context: Léopold Eyharts (born April 28, 1957) is a Brigadier General in the French Air Force, an engineer and ESA astronaut. Prof. Dr. Ulrich Hans Walter (born February 9, 1954) is a German physicist/engineer and a former DFVLR astronaut.
Substitutes: French, German

(b)
Question: Are the movies "Monsters, Inc." and "Mary Poppins" both by the same company?
Context: Mary Poppins is a 1964 American musical-fantasy film directed by Robert Stevenson and produced by Walt Disney, with songs written and composed by the Sherman Brothers. Monsters, Inc. is a 2001 American computer-animated comedy film produced by Pixar Animation Studios and distributed by Walt Disney Pictures.
Substitutes: Walt Disney, Universal

(c)
Question: What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?
Context: Peter Bolesław Schmeichel MBE (born 18 November 1963) is a Danish former professional footballer who was voted the IFFHS World's Best Goalkeeper in 1992 and 1993. Kasper Peter Schmeichel (born 5 November 1986) is a Danish professional footballer. He is the son of former Manchester United and Danish international goalkeeper Manuel Neuer.
AdvSent1: Robert Lewandowski was voted to be the World's Best Striker in 1992.
AdvSent2: Michael Jordan was voted the IFFHS best NBA player in 1992.

Figure 4: Examples (contexts are truncated for brevity) of our property annotations on HotpotQA base data points. The top two are yes/no questions and the third is a bridge question.

Approach	Yes-No		Bridge	
	S-ACC	S-AUC	S-ACC	S-AUC
MAJORITY	52.0	–	56.0	–
CONF	64.0	49.8†	66.0	65.9†
LIME	72.0	73.6†	74.0	71.4†
SHAP	72.0	70.5†	76.0	75.0
INTGRAD	72.0	75.2†	72.0	77.9
DIFFMASK	66.0	60.2†	68.0	62.3†
ARCHIP	56.0	53.2†	62.0	57.5†
ATATTR	66.0	63.6†	72.0	79.1
LATATTR	**84.0**	**87.9**	**78.0**	**81.7**

Table 1: Results on HOTPOTQA Yes-No type and Bridge questions. We perform significance tests on accuracy via bootstrap resampling for the comparisons between LATATTR and other approaches. A dagger indicates a method for which our approach outperforms it by a statistically significant margin ($p < 0.05$). Overall, the best pairwise technique yields better simulation accuracy than the best token-level technique.

77.2 F1 on the development set in the distractor setting, comparable to other strong ROBERTA-based models (Tu et al., 2020; Groeneveld et al., 2020).

In the SQUAD-ADV setting, we also use a ROBERTA QA model which achieves 92.2 F1 on the SQUAD dev set and 68.0 F1 on SQUAD-ADV. Our model is trained on SQUAD v1.0 for 4 epochs using a learning rate of 1.5e-5, a batch size of 32 and a warm-up ratio of 6%.

5.1 Hotpot Yes-No Questions

We first study a subset of yes/no comparison questions, which are challenging despite the binary answer space (Clark et al., 2019). Typically, a yes-no comparison type question requires comparing the properties of two entities (Figure 2).

Hypothesis & Counterfactuals The hypothesis $\mathcal{H}$ we investigate is as in Section 2: *the model compares the entities' properties as indicated by the question.* Most Hotpot Yes-No questions follow one of two templates: *Are A and B both __?* (Figure 4a), and *Are A and B of the same __?* (Figure 4b). We define the **property** tokens associated with each question as the tokens *in the context* that match the blank in the template; that is, the values of the property that A and B are being compared on. For example, in Figure 4a, *French* and *German* are the property tokens, as the property of interest is the national origin.

To construct a neighborhood for a base data point, we take the following steps: (1) manually extract the property tokens in the context; (2) replace the property token with two substitutes, forming a set of four counterfactuals exhibiting nonidentical ground truths. When the properties associated with the two entities differ from each other, we directly use the properties extracted as the substitutes (Figure 4a); otherwise we add a new property candidate that is of the same class (Figure 4b).

We set $z = 0$ (the hypothesis does not hold) if for each perturbed example $D_i \in \mathcal{D}$, the model predicts the same answer as for the original example, indicating a failure to compare the properties. We set $z = 1$ if the model's prediction *does* change. We choose a binary scheme to label the model behavior because we observed that, on the small perturbation sets, the model performance was bimodal: either the tokens mattered for the prediction (reflected by the model changing its prediction at least once) or they didn't. The authors annotated perturbations for 50 ($\mathcal{D}$, z) randomly selected pairs in total, forming a total of 200 counterfactual instances. Full counterfactual sets are available with our data and code release.

Connecting Explanation and Hypothesis To make a judgment about z, we extract a factor f based on the importance of a set of property tokens

P. For token attribution-based methods, we define f as the sum of the attribution s_i of each token in P: $\sum_{i \in P} s_i$. For feature interaction-based methods producing pairwise attribution s_{ij}, we compute f by pooling the scores of all the interaction related to the property tokens: $\sum_{i \in P \vee j \in P} s_{ij}$.

Now we predict $z = 1$ if the factor f is above a threshold, and evaluate the capability of the factor in indicating the model high-level behavior using the best simulation accuracy it can achieve (S-ACC) and AUC score (S-AUC).[6]

Results First, we show that using attributions can indeed help predict the model's behavior. In Table 1, our approach (LATATTR) is the best, achieving a simulation accuracy of 84%. That is, with a properly set threshold, we can successfully predict whether the model predictions change when perturbing the properties in the original example 84% of the time. The attributions therefore give us the ability to simulate our model's behavior better than the other methods here. Our approach also improves substantially over the vanilla ATATTR method.

The best token-level attribution based approaches obtain an accuracy of 72%, significantly lagging the best pairwise technique. This indicates token attribution based methods are less effective in the HOTPOTQA Yes-No setting; we hypothesize that this is due to the importance of token interaction in this RC setting.

In this setting, DIFFMASK performs poorly, typically because it assigns high attribution to many tokens, since it determines which tokens need to be kept rather than distinguishing fine-grained importance (see the appendix for examples). It's possible that other heuristics or models learned on large numbers of perturbations could more meaningfully extract predictions from this technique.

5.2 Hotpot Bridge Questions

We also evaluate the explanation approaches on so-called bridge questions on the HOTPOTQA dataset, described in Yang et al. (2018). Figure 5 shows an example explanation of a bridge question. From the attribution scores we find the most salient connection is between the span "*what government position*" in the question and the span "*United States Ambassador*" in the context. This attribution directly highlights the reasoning shortcut (Jia and Liang, 2017; Chen and Durrett, 2019; Min et al., 2019; Jiang and Bansal, 2019) the model is using, where it disregards the second part of the question. If we inject an additional sentence *"Hillary Clinton is an American politician, who served as the United States secretary of the state from 2009 to 2013"* into the context, the model will be misled and predict "*United States secretary*" as the new answer. This sentence could easily have been part of another document retrieved in the retrieval stage, so we consider its inclusion to be a realistic counterfactual.

[6]Note that for different attribution methods, the thresholds are different and set to achieve the best accuracy.

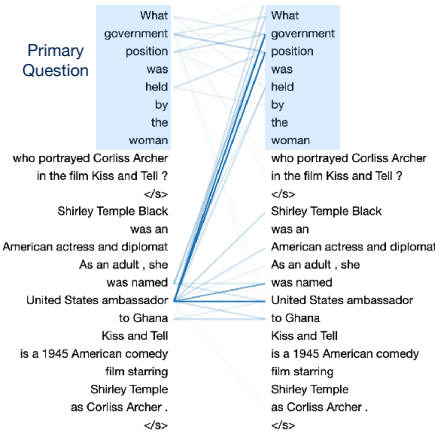


Figure 5: Explanations generated by our approach for a bridge question from HOTPOTQA. The prediction can mostly be attributed to the primary question, indicating the model is taking a reasoning shortcut, and the prediction is flipped with an adversarial attack.

We further define the *primary* question, i.e., the span of the question containing wh-words with heavy modifier phrases and embedded clauses dropped, following the decomposition principle from Min et al. (2019). In Figure 5), the primary question is "*What government position is held by the woman*."

Hypothesis & Counterfactuals The hypothesis $\mathcal{H}$ we investigate is: *the model is using correct reasoning and not a shortcut driven by the primary question*.

We construct counterfactuals following the same idea applied in our example. We view bridge questions as consisting of two single hop questions, the primary part and the secondary part. For a given question, we add an adversarial sentence based on the primary question so as to alter the model prediction. The added adversarial sentence contains context leading to a spurious answer to only the primary question, but does not change the gold answer.

We do this twice, yielding a set $\mathcal{D} = \{D_0, D_1, D_2\}$ consisting of the base example and two perturbations. We define the label of D to be $z = 0$ in the case that model's prediction does change under the perturbations, and $z = 1$ otherwise. We show one example in Figure 4c. More examples and the full counterfactual set can be found in the appendix.

We randomly sample 50 base data points from the development set and two authors each write an adversarial sentence, giving 150 data points total.

Connecting Explanation and Hypothesis For this setting, we use a factor describing the importance of the primary question normalized by the importance of the entire question. Namely, let $P = \{p_i\}$ be the set of tokens in the primary questions, and $Q = \{q_i\}$ be the set of tokens in the entire question. We define the factor f as the importance of P normalized by the importance of Q, where the importance calculation is the same as in Section 5.1. A higher factor means it is more heavily relying only on the primary question and hence a better chance of being attacked.

Results According to the simulation ACC scores in Table 1, token-level attributions are somewhat more successful at indicating model behavior in this setting compared to the yes/no setting. Our approach as the best feature interaction based technique is able to achieve a stimulation accuracy of 78%, slightly outperforming the best token attribution approach.

5.3 SQuAD Adversarial

Hypothesis & Counterfactuals Our hypothesis $\mathcal{H}$ is: *the model can resist adversarial attacks of the addSent variety* (Jia and Liang, 2017). For each of the original examples D_0 from a portion of the SQUAD-ADV development set, Jia and Liang (2017) creates 5 adversarial attacks, which are paraphrased and filtered by Turkers to give 0 to 5 valid attacks for each example, yielding our set $\mathcal{D}$. We define the label of $\mathcal{D}$ to be $z = 1$ if the model resists all the adversarial attacks posed on D_0 (i.e., predictions for D are the same). To ensure the behavior is more precisely profiled by the counterfactuals, we only keep the base examples with more than 3 valid attacks, resulting in a total number of 276 $(\mathcal{D}, z)$ pair (1,506 data points).

Connecting Explanation and Hypothesis We use a factor f indicating the importance of the essential keywords extracted from the question using POS tags (proper nouns and numbers). E.g., for the question "*What Florida stadium was considered for Super Bowl 50*", we extract "*Florida*", "*Super Bowl*" , and "*50*". If the model considers all the essential keywords mentioned in the question, it should not be fooled by distractors with irrelevant information. We show a set of illustrative examples in the appendix. We compute the importance scores in the same way described in Section 5.1.

In addition to the scores provided by various explanation techniques, we also use the model's confidence on the original prediction as a baseline.

Approach	S-ACC	S-AUC
MAJORITY	52.1	–
CONF	58.3	57.8†
LIME	67.7	68.3†
SHAP	65.9	68.3†
INTGRAD	61.6	61.1†
DIFFMASK	57.6	53.6†
ARCHIP	58.6	56.2†
ATATTR	69.4	**72.5**
LATATTR	**70.0**	72.1

Table 2: Simulation Accuracy and AUC scores for the SQuAD adversarial setting, assessing whether the model changes its prediction on an example when attacked. We perform significance tests on accuracy via bootstrap resampling for the comparisons between LATATTR and other approaches. A dagger indicates a method for which our approach outperforms it by a statistically significant margin ($p < 0.05$).

Results We show results in Table 2. The best approaches (ATATTR and LATATTR) can achieve a simulation accuracy around 70%, 10% above the performance based on raw model confidence. This shows the model is indeed over-confident in its predictions; our assumption about the robustness together with our technique can successfully expose the vulnerability in some of the model predictions.

There is room to improve on these results; our simple heuristic cannot perfectly connect the explanations to the model behavior in all cases. We note that there are other orthogonal approaches (Kamath et al., 2020) to calibrate the confidence of QA models' predictions by looking at statistics of the adversarial examples. Because our goal is to assess attributions rather than optimize for calibration, our judgment is made purely based on the *original* example, and does not exploit learning to refine our heuristic.

5.4 Discussion and Limitations

We show that feature attributions can reveal known dataset biases and reasoning shortcuts in HotpotQA without having to perform a detailed manual analysis. This confirms the suitability of our attribution methods for at least this use case: model designers can look at them in a semi-automated way and determine how robust the model is going to be when faced with counterfactuals.

Our analysis also highlights the limitations of current explanation techniques. We experimented with other counterfactuals by permuting the order of the paragraphs in the context, which often gave rise to different predictions. We believe the model prediction was in these cases impacted by biases in positional embeddings (e.g., the answer tends to occur in the first retrieved paragraph), which cannot be indicated by current attribution methods. We believe this is a useful avenue for future investigation. By first thinking about what kind of counterfactuals and what kind of behaviours we want to explain, we can motivate the development of new explanation techniques to serve these needs.

6 Related Work

We focus on several prominent token attribution techniques, but there are other related methods as well, including other methods based on Shapley values (Štrumbelj and Kononenko, 2014; Lundberg and Lee, 2017), contextual decomposition (Jin et al., 2020), and hierarchical explanations (Chen et al., 2020). These formats can also be evaluated using our framework if being connected with model behavior using the proper heuristic. Other work explores so-called concept-based explanations (Mu and Andreas, 2020; Bau et al., 2017; Yeh et al., 2019). These provide another pathway towards building explanations of high-level behavior; however, they have been explored primarily for image recognition tasks and cannot be directly applied to QA, where defining these sorts of concepts is challenging. Finally, textual explanations (Hendricks et al., 2016) are another popular alternative, but it is difficult to evaluate these in our framework as it is very difficult to bridge from a free-text explanation to an approximation of a model's computation.

Probing techniques aim to discover what intermediate representations have been learned in neural models (Tenney et al., 2019; Conneau et al., 2018; Hewitt and Liang, 2019; Voita and Titov, 2020). Internal representations could potentially be used to predict behavior on contrast sets similar to this work; however, this cannot be done heuristically and larger datasets are needed to explore this.

Other work considering how to evaluate explanations is primarily based on how explanations can assist humans in predicting model decisions for a given example (Doshi-Velez and Kim, 2017; Chandrasekaran et al., 2018; Nguyen, 2018; Hase and Bansal, 2020); We are the first to consider building contrast sets for this. Similar ideas have been used in other contexts (Kaushik et al., 2020; Gardner et al., 2020) but our work focuses on evaluation of explanations rather than general model evaluation.

7 Conclusion

We have presented a new methodology using explanations to understand model behavior on realistic counterfactuals. We show explanations can indeed be connected to model behavior, and therefore we can compare explanations to understand which ones truly give us actionable insights about what our models are doing.

We have showcased how to apply our methodology on several RC tasks, leveraging either semi-automatically curated counterfactual sets or existing resources. We generally find pairwise interaction methods perform better than the best token-level attribution based methods in our analysis. More broadly, we see our methodology as a useful evaluation paradigm that could be extended across a range of tasks, leveraging either existing contrast sets or with a small amount of effort devoted to create targeted counterfactual sets as in this work.

Acknowledgments

Thanks to Eunsol Choi, Jiacheng Xu, Jifan Chen, Qiaochu Chen, and everyone in the UT TAUR lab for helpful discussions, as well as to the anonymous reviewers for their helpful feedback. This work was partially supported by NSF Grant IIS-1814522, a gift from Arm, a gift from Salesforce Inc, and an equipment grant from NVIDIA.

References

Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. 2018. Verifiable reinforcement learning via

policy extraction. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Dirk Groeneveld, Tushar Khot, Mausam, and Ashish Sabharwal. 2020. A simple yet strong pipeline for HotpotQA. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in NLP. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

David Harbecke and Christoph Alt. 2020. Considering likelihood in NLP classification explanations with occlusion and language modeling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating Visual Explanations. In *European Conference on Computer Vision (ECCV)*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

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

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

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

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, abs/1907.11692.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1.

Soumya Sanyal and Xiang Ren. 2021. Discretized Integrated Gradients for Explaining Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do Human Rationales Improve Machine Explanations? In *Proceedings of the ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP)*.

Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.

Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. 2020. Obtaining faithful interpretations from compositional neural networks. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Generating token-level explanations for natural language inference. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. 2020. How does this interaction affect me? interpretable attribution for feature interactions. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. 2018. Programmatically interpretable reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP Demo Track)*.

Jialin Wu and Raymond Mooney. 2019. Faithful multimodal explanation for visual question answering. In *Proceedings of the ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP)*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chih-Kuan Yeh, Been Kim, Sercan Ö. Arik, C. Li, P. Ravikumar, and T. Pfister. 2019. On concept-based explanations in deep neural networks. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

A Details of Hotpot Yes-No Counterfactuals

Figure 6 shows several more examples to illustrate our process of generating counterfactuals for the Hotpot Yes-No setting.

As stated in Section 5.1, most Hotpot Yes-No questions follow one of two templates: *Are A and B both __?* (Figure 6, abc), and *Are A and B of the same __?* (Figure 6, def). The property tokens that match the blank in the template are highlighted in Figure 6.

Recall our two steps to construct a neighborhood for a base data point:

1. Manually extract the property tokens in the context
2. Replace each property token with a substitute, forming a set of four counterfactuals exhibiting nonidentical ground truths

When the properties associated with the two entities differ from each other, we directly use the properties extracted as the substitutes (Figure 6, abf); otherwise we add a new property candidate that is of the same class (Figure 6, cde).

We annotated randomly sampled examples from the Hotpot Yes-No questions. We skipped several examples that compared abstract concepts with no explicit property tokens. For instance, we skipped the question *Are both Yangzhou and Jiangyan District considered coastal cities?* whose given context does not explicitly mention whether the cities are coastal cities. We looked through 61 examples in total and obtained annotations for 50 examples, so such discarded examples constitute a relatively small fraction of the dataset. Overall, this resulted in 200 counterfactual instances. We found the prediction of a RoBERTa QA model on 52% of the base data points change when being perturbed.

B Details of Hotpot Bridge Counterfactuals

Figure 7 shows more examples of our annotations for generating counterfactuals for Hotpot Bridge examples. We first decompose the bridge questions into two single hop questions (Min et al., 2019), the **primary** part (marked in Figure 7) and **secondary** part. The primary part is the main body of the question, whereas the secondary part is usually a clause used to link the bridge entity (Min et al., 2019).

Next, we write adversarial sentences for confuse the model follow a similar method used for generating SQuAD adversarial examples (Jia and Liang, 2017). Specifically, we only look at the primary part, and write down a sentence that can answer the primary question accordingly with a different entity from the secondary question. This will introduce a spurious answer, but does not change the gold answer. Besides, we also write the sentences follow in the same Wikipedia style as the original context possible, and some of the sentences are modified from texts from Wikipedia (e.g., Figure 7 ac).

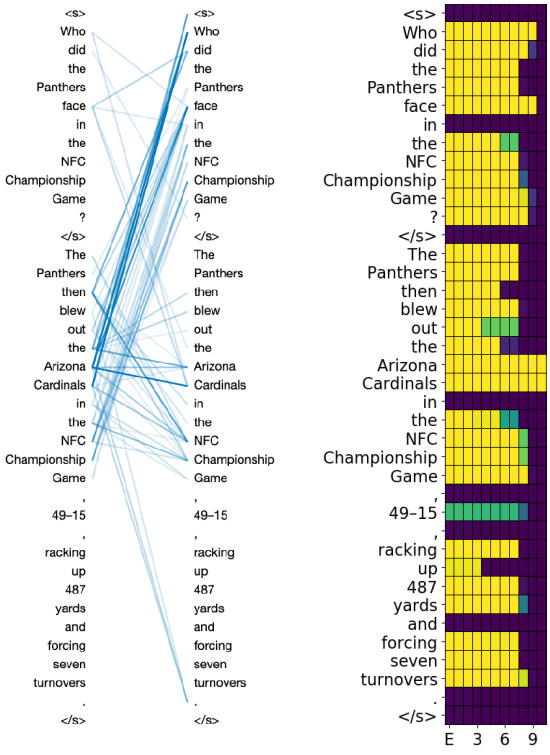
Two of the authors each wrote a single adversarial sentence for 50 of the Hotpot Bridge examples, yielding 150 counterfactual instances in total. The adversarial sentences manage to alter 56% of the predictions on the base examples.

(a)	Question	Were Ulrich Walter and Léopold Eyharts both from Germany?
	Context	Léopold Eyharts (born April 28, 1957) is a Brigadier General in the French Air Force, an engineer and ESA astronaut. Prof. Dr. Ulrich Hans Walter (born February 9, 1954) is a German physicist/engineer and a former DFVLR astronaut.
	Substitutes	French, German
(b)	Question	Are both Aloinopsis and Eriogonum ice plants?
	Context	Aloinopsis is a genus of ice plants from South Africa. Eriogonum is the scientific name for a genus of flowering plants in the family Polygonaceae. The genus is found in North America and is known as wild buckwheat.
	Substitutes	ice, flowering
(c)	Question	Were Frank R. Strayer and Krzysztof Kieślowski both Directors?
	Context	Frank R. Strayer (September 21, 1891 - 2013 February 3, 1964) was an actor, film writer, and director. He was active from the mid-1920s until the early 1950s. Krzysztof Kieślowski (27 June 1941 - 13 March 1996) was a Polish art-house film director and screenwriter.
	Substitutes	director, producer
(d)	Question	Were Scott Derrickson and Ed Wood of the same nationality?
	Context	Scott Derrickson (born July 16, 1966) is an American director, screenwriter and producer. Edward Davis Wood Jr. (October 10, 1924 - 2013 December 10, 1978) was an American filmmaker, actor, writer, producer, and director.
	Substitutes	American, English
(e)	Question	Are the movies "Monsters, Inc." and "Mary Poppins" both by the same company?
	Context	Mary Poppins is a 1964 American musical-fantasy film directed by Robert Stevenson and produced by Walt Disney, with songs written and composed by the Sherman Brothers. Monsters, Inc. is a 2001 American computer-animated comedy film produced by Pixar Animation Studios and distributed by Walt Disney Pictures.
	Substitutes	Walt Disney, Universal
(f)	Question	Are Steve Perry and Dennis Lyxzén both members of the same band?
	Context	Stephen Ray Perry (born January 22, 1949) is an American singer, songwriter and record producer. He is best known as the lead singer of the rock band Journey. Dennis Lyxzén (born June 19, 1972) is a musician best known as the lead vocalist for Swedish hardcore punk band Refused.
	Substitutes	Journey, Refused

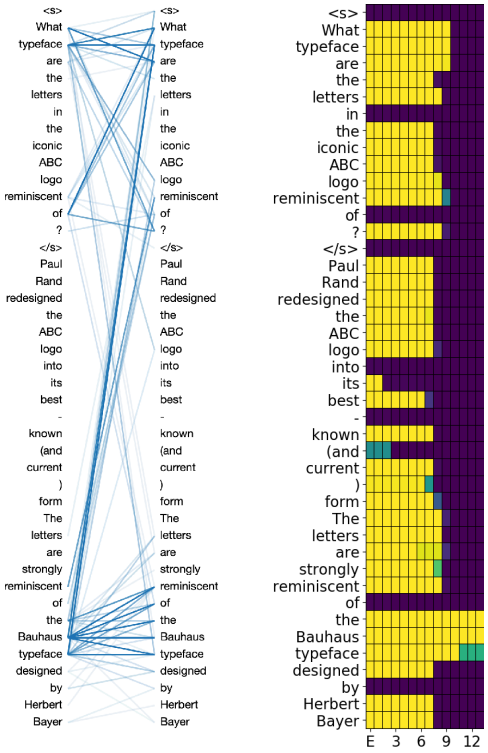
Figure 6: Examples (contexts are truncated for brevity) of our annotations on Hotpot Yes-No base data points. We find the property tokens in the context, and build realist counterfactuals by replacing them with substitutes that are properties extracted in the base data point or similar properties hand-selected by us.

(a)	Question	What is the name of the fight song of the university whose main campus is in Lawrence, Kansas and whose branch campuses are in the Kansas City metropolitan area?
	Context	Kansas Song (We're From Kansas) is a fight song of the University of Kansas. The University of Kansas, often referred to as KU or Kansas, is a public research university in the U.S. state of Kansas. The main campus in Lawrence, one of the largest college towns in Kansas, is on Mount Oread, the highest elevation in Lawrence. Two branch campuses are in the Kansas City metropolitan area.
	Adv Sent 1	Texas Fight is a fight song of the University of Texas at Austin.
	Adv Sent 2	Big C is a fight song of the University of California, Berkeley.
(b)	Question	What screenwriter with credits for "Evolution" co-wrote a film starring Nicolas Cage and Téa Leoni?
	Context	David Weissman is a screenwriter and director. His film credits include "The Family Man" (2000), "Evolution" (2001), and "When in Rome" (2010). The Family Man is a 2000 American romantic comedy-drama film directed by Brett Ratner, written by David Diamond and David Weissman, and starring Nicolas Cage and Téa Leoni.
	Adv Sent 1	Don Jakoby is an American screenwriter that collabrates with David Weissman in "Evolution".
	Adv Sent 2	Damien Chazelle is a screenwriter most notably known for writing La La Land.
(c)	Question	The arena where the Lewiston Maineiacs played their home games can seat how many people ?
	Context	The Androscoggin Bank Colisée (formerly Central Maine Civic Center and Lewiston Colisee) is a 4,000 capacity (3,677 seated) multi-purpose arena, in Lewiston, Maine, that opened in 1958. The Lewiston Maineiacs were a junior ice hockey team of the Quebec Major Junior Hockey League based in Lewiston, Maine. The team played its home games at the Androscoggin Bank Colisée.
	Adv Sent 1	Allianz (known as Fußball Arena München for UEFA competitions) is a arena in Munich, with a 5,000 seating capacity.
	Adv Sent 2	The Tacoma Dome is a multi-purpose arena (221,000 capacity, 10,000 seated) in Tacoma, Washington, United States.
(d)	Question	Scott Parkin has been a vocal critic of Exxonmobil and another corporation that has operations in how many countries ?
	Context	Scott Parkin (born 1969, Garland, Texas is an anti-war, environmental and global justice organizer, former community college history instructor, and a founding member of the Houston Global Awareness Collective. He has been a vocal critic of the American invasion of Iraq, and of corporations such as Exxonmobil and Halliburton. The Halliburton Company, an American multinational corporation. One of the world's largest oil field service companies, it has operations in more than 70 countries.
	Adv Sent 1	Visa is a corporation that has operations in more than 200 countries.
	Adv Sent 2	The Ford Motor Company is an American multinational corporation with operations in more than 100 countries.
(e)	Question	In 1991 Euromarché was bought by a chain that operated how any hypermarkets at the end of 2016?
	Context	Carrefour S.A. is a French multinational retailer headquartered in Boulogne Billancourt, France, in the Hauts-de-Seine Department near Paris. It is one of the largest hypermarket chains in the world (with 1,462 hypermarkets at the end of 2016). Euromarché was a French hypermarket chain. In June 1991, the group was rebought by its rival, Carrefour, for 5,2 billion francs.
	Adv Sent 1	Walmart Inc is a multinational retail corporation that operates a chain of hypermarkets that owns 4,700 hypermarkets within the United States at the end of 2016.
	Adv Sent 2	Trader Joe's is an American chain of grocery stores headquartered in Monrovia, California. By the end of 2016, Trader Joe's had over 503 stores nationwide in 42 states.
(f)	Question	What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992?
	Context	Peter Bolesław Schmeichel MBE (born 18 November 1963) is a Danish former professional footballer who played as a goalkeeper, and was voted the IFFHS World's Best Goalkeeper in 1992 and 1993. Kasper Peter Schmeichel (born 5 November 1986) is a Danish professional footballer. He is the son of former Manchester United and Danish international goalkeeper Manuel Neuer.
	Adv Sent 1	Robert Lewandowski was voted to be the World's Best Striker in 1992.
	Adv Sent 2	Michael Jordan was voted the IFFHS best NBA player in 1992.

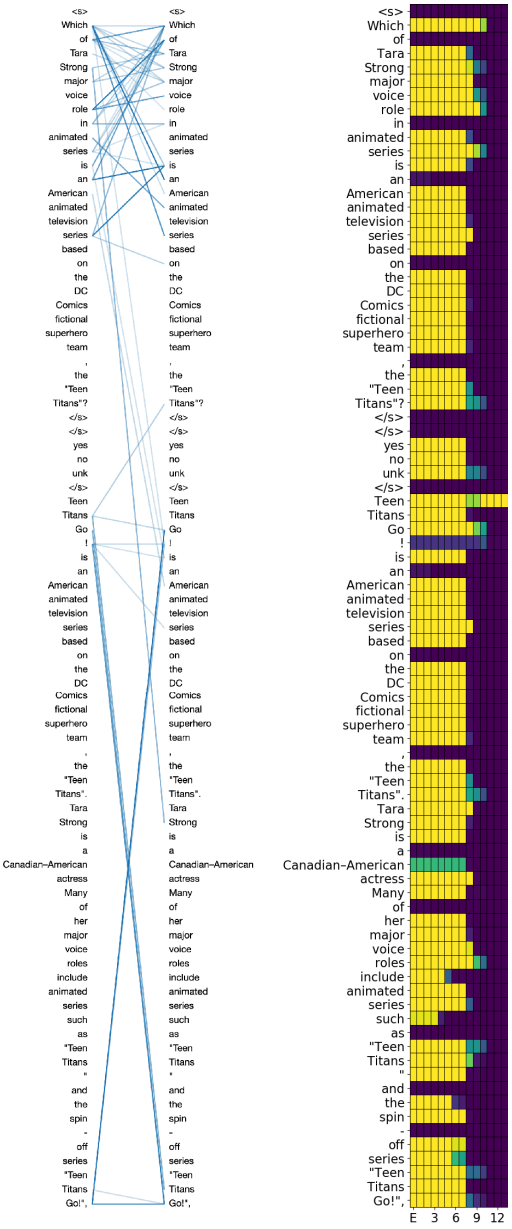
Figure 7: Examples (contexts are truncated for brevity) of primary questions and adversarial senteces for creating Hotpot Bridge counterfactuals.



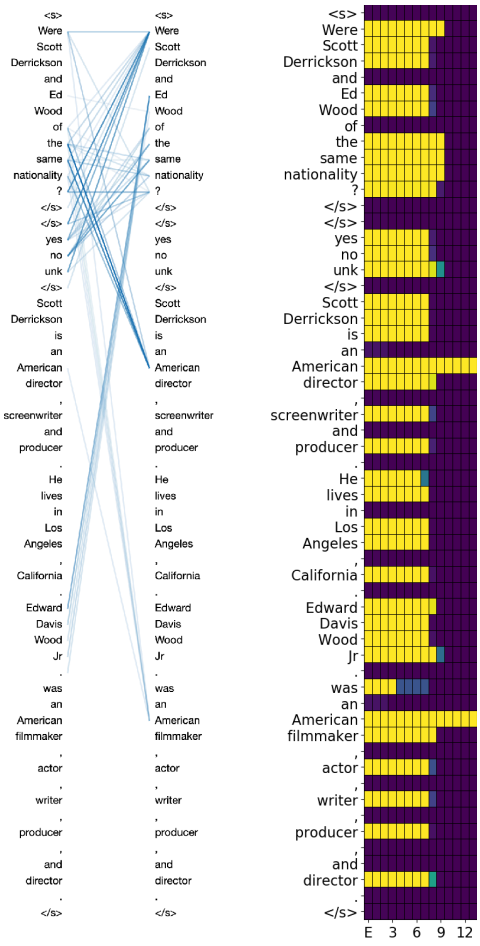
(a) Explanations generated by our approach (left) and DIFFMASK (right) for an example from SQUAD dataset. We automatically extract "Panthers" and "NFC Championship Game" as the essential keywords. From the explanation, we see these keywords contribute to the model prediction, and therefore infer the model is more likely to be able to resist adversarial attacks posed on this example.



(b) Explanations generated by our approach (left) and DIFFMASK (right) for an example from SQUAD dataset. When adding an adversarial sentence, the model changes its prediction. Our explanation clearly shows the model bases its prediction on "what typeface" without taking into account the entity "ABC".



(a) Explanations generated by our approach (left) and DIFFMASK (right) for a bridge example from the HOTPOTQA dataset. The model can resist adversarial sentences posed on this example. Our explanation highlights certain tokens in the latter part of the question ("American animated television" and "Teen Titans"), suggesting the model prediction is less likely to be flipped by adversarial attacks targeted at this example, which aligns with the model behavior.



(b) Explanations generated by our approach (left) and DIFFMASK (right) for a comparison example from the HOTPOTQA dataset. When we perturb the nationalities in the context, the model changes its prediction. Both of the explanations both suggest the model makes its decision by looking at the nationalities associated with the two entities, which is congruent with the model behavior.