Logical Satisfiability of Counterfactuals for Faithful Explanations in NLI

Suzanna Sia[1*], **Anton Belyy**[1], **Amjad Almahairi**[2],
Madian Khabsa[2], **Luke Zettlemoyer**[2], **Lambert Mathias**[2]

[1] Johns Hopkins University
[2] Meta AI Research

Abstract

Evaluating an explanation's faithfulness is desired for many reasons such as trust, interpretability and diagnosing the sources of model's errors. In this work, which focuses on the NLI task, we introduce the methodology of Faithfulness-through-Counterfactuals, which first generates a counterfactual hypothesis based on the logical predicates expressed in the explanation, and then evaluates if the model's prediction on the counterfactual is consistent with that expressed logic (i.e. if the new formula is logically satisfiable). In contrast to existing approaches, this does not require any explanations for training a separate verification model. We first validate the efficacy of automatic counterfactual hypothesis generation, leveraging on the few-shot priming paradigm. Next, we show that our proposed metric distinguishes between human-model agreement and disagreement on new counterfactual input. In addition, we conduct a sensitivity analysis to validate that our metric is sensitive to unfaithful explanations.

1 Introduction

How should we evaluate an explanation's *faithfulness* with respect to the task model? According to Jacovi and Goldberg (2020), faithful measures should focus on utility to the user and the idea that an explanation can be *sufficiently faithful*.[1] The goal of interpretability research is to increase *warranted* trust and identify the influence of certain variables and allow users to understand how a model will behave on given inputs (Doshi-Velez and Kim 2017; Lipton 2018).

In interpretable NLP, there is growing interest in tasks that require world and commonsense "knowledge" and "reasoning" (Danilevsky et al. 2020). We focus on natural language inference (SNLI; Bowman et al. (2015)), where extractive explanations also known as rationales (DeYoung et al. 2020) are limited as they take a subset of the existing input. Instead, we require *free-form natural language explanations* to fill in the reasoning or knowledge gap for such tasks (Camburu et al. 2018; Rajani et al. 2019). Our setting is thus characterised by the post-hoc interpretation of black-box classification models via generative explanations. Our work follows the standard "predict-and-explain" paradigm (Do et al. 2020). Here an explanation generator generates the explanations conditioned on the predicted task label.[2] Without faithfulness evaluations, the explanation approximately describes the internal process at best, and is generated from superficial similarities between the training data and the class label at worst.

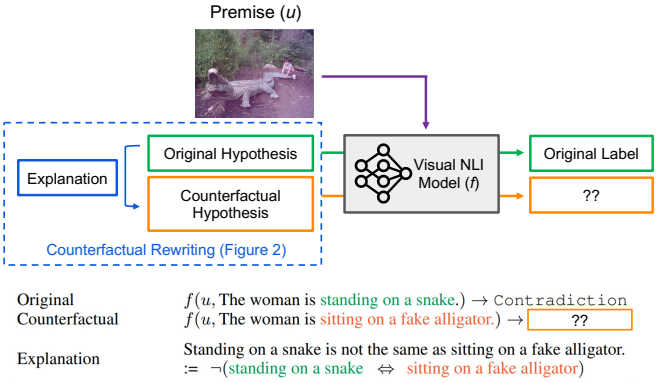

Figure 1: Overview of the proposed FTC approach, evaluating **f**aithfulness of explanations **t**hrough **c**ounterfactuals. If the explanation is faithful to the model, the NLI label on the new counterfactual hypothesis should change to `Entailment`. If the model still predicts `Contradiction`, this indicates that the explanation is not faithful to the model, i.e. the logic of the explanation and the model are not consistent.

The central contribution of this paper is a **methodology grounded in first-order free logic (Lambert 1967; Bencivenga 2002),[3] to verify a given explanations' faithfulness**. Our proposed approach generates a revised (counterfactual) hypothesis based on the logical propositions expressed in the explanation, and evaluates the *logical satisfiability* (Boolos,

[*]Work done at Meta internship. Correspondence to ssia1@jh.edu

[1]Jacovi and Goldberg (2020) originally posit that faithful explanations should "accurately represents the reasoning process behind the model's prediction", however also acknowledge that this is "impossible to satisfy fully".

[2](Do et al. 2020) report little to no difference between jointly predict and explain compared to predict then explain. Note that the emphasis of this work is on evaluating the faithfulness of explanations rather than generating them.

[3]This is an extension of predicate logic and should not be confused with either predicate or propositional logic.

Burgess, and Jeffrey 2002) of the new hypothesis. Consider the following example:

Hypothesis:	The dog is barking at the girl.
Explanation:	The dog is an animal.
Counterfactual:	The animal is barking at the girl.

If the explanation is logically consistent (faithful) to the model, then the revised counterfactual hypothesis which replaces 'dog' with 'animal' in the original hypothesis, should be satisfiable, since 'dog is an animal'. However, if the explanation is inconsistent with the model, then the resulting hypothesis is unsatisfiable and the explanation is unfaithful. We describe this formally in Section 2.

Compared to previous automatic metrics (LAS; Hase et al. (2020), LRA; Wiegreffe, Marasović, and Smith (2021)), our proposed method does not rely on an external verification model and does not require explanation data for training.[4] Our method directly queries the task model in question while crucially avoiding the confound of "label leakage"[5] from the explanation (Hase et al. 2020). We expand on this discussion in Section 4.

The contributions of this work are as follows:

- We present a methodology for evaluating faithfulness of free form explanations for NLI, grounded in first-order free logic (Section 2.1, Section 2.2). Our method evaluates the satisfiability of logical relations expressed through a counterfactual hypothesis.
- We introduce an automatic metric (Section 2.3) and show its viability with human studies, indicating that a practical solution exists for the proposed (theoretical) method. We leverage few-shot priming for generating counterfactual hypothesis, achieving $0.71 - 0.88$ METEOR score for human and generated explanations (Section 3.2).
- We further evaluate the effectiveness of the fully automated approach for the proposed metric in distinguishing when the model's prediction agrees and disagrees with gold labels on gold counterfactual hypothesis, and achieve $0.58 - 0.75$ ρ-statistic on Wilcoxon rank-sum test (Section 3.3). Our method is sensitive to pathological explanations that were generated by removing inputs to the explanation generator, as compared to other existing faithfulness metrics (Section 3.4).

2 Method

2.1 Problem Formulation

Natural Language Inference (NLI) is typically cast as a classification task; given a premise u and a hypothesis x, the classifier f predicts the label y, where $y \in \{\texttt{E}, \texttt{C}, \texttt{N}\}$. Here E indicates entailment, C contradiction, and N neutral, for the relationship between u and x. f can therefore be viewed as a black-box function approximating the solution to a *logical satisfiability* problem.

Label	Propositions	Description
E	$u \Rightarrow x$	hypothesis implied by premise
C	$u \Rightarrow \neg x$	hypothesis contradicts premise
N	$u \overset{(?)}{\Longrightarrow} x$	hypothesis neither contradicts or is entailed by premise

Table 1: Mapping NLI task labels to propositions. $\Rightarrow$ indicates logical implication, $\neg$ indicates logical negation, and $\overset{(?)}{\Longrightarrow}$ indicates *truth-valueless*.

Testing Predicate Relations in Explanations. An explanation z can express one or more logical predicate relations (R), which describes the relationship between two variables A and B that are expressed in x and u respectively (see Table 2 col 4 for examples). This is denoted as $R(A, B)$.[6]

A "faithful" explanation with respect to a task model f, is one that expresses predicate relations $R(A, B)$ that are *consistent* with f's predictions. The central idea of this work, is to automatically verify this using a *counterfactual hypothesis*, x^{cf}, and its derived associated *counterfactual label* (the expected satisfiability result). The *what-if* question is, if $f(u, do(x = x^{\text{cf}}))$ does not result in the associated counterfactual label, then the explanation is not faithful to f. In the language of counterfactuals, the do-operator forces a variable to take a certain value or distribution (Pearl 1995). We write $do(x = x^{\text{cf}})$ as x^{cf} for notation simplicity.

First-order Free Logic in NLI. In order to deduce the associated label for x^{cf}, we must address the issue of logical deduction for Neutral, which has no corresponding expression in classical predicate logic. We observe that the ternary label in NLI parallels first-order free logic, which has three distinct logical forms, positive, negative and neutral (Lambert 1967; Nolt 2021). In contrast to classical logic which requires each singular term to denote a Boolean variable in the domain, free logic may have formula which are *truth-valueless* (Nolt 2021), i.e., it is not known whether they are True or False.[7] Table 1 shows the task label, propositions and their meaning in Free Logic.

2.2 Satisfiabilty of the Counterfactual

In this section, we derive what the counterfactual hypothesis and associated counterfactual label would be for each original label assuming logical formulas and discrete variables where exact substitution is possible. Section 2.4 describes our suggested approach to handle Natural Language where discrete substitution no longer holds.

[4] "Faithfulness" measures which are tied to an external verification model are potentially problematic as given a fixed task model and explanation, one could in theory achieve two different faithfulness scores if the verification model changes.

[5] Label leakage occurs because of superficial similarities between the syntatic form of the explanation and the task label. For instance the explanations "A is a B" and "A is not a B" are highly associated with the Entailment and Contradiction label.

[6] While a typical presentation of R involves more semantically meaningful such as "Father(A, B), relations involving only the logical operators are considered predicate relations.

[7] Truth-valueless formulas are often said to have "truth-value gaps". Informally, this can be interpreted as there being insufficient information on the truth-values of logical variables to conclude the relationship between u and x (Nolt 2021).

Label (y)	Propositional Formula	Explanation (z)	$R(A, B)$	Propositional Formula (cf)	Label (y^{cf})
E	$u \Rightarrow x$	A is the same as B	$A \Leftrightarrow B$	$u \Rightarrow x^{\text{cf}}$	E
C	$u \Rightarrow \neg x$	A is not B	$\neg(A \Leftrightarrow B)$	$u \Rightarrow \neg(\neg x^{\text{cf}})$	E
N	$u \overset{(?)}{\Longrightarrow} x$	A does not imply B	$\neg(A \Rightarrow B)$	$u \Rightarrow x^{\text{cf}}_{[A]}$	E
				$u \overset{(?)}{\Longrightarrow} x^{\text{cf}}_{[B]}$	N

Table 2: From the original hypothesis x and logical predicate relations $R(A, B)$ expressed in the explanation, we generate the counterfactual hypothesis x^{cf} (2 cases $x^{\text{cf}}_{[A]}$ and $x^{\text{cf}}_{[B]}$ for $y = \texttt{N}$). The resulting counterfactual label y^{cf} is logically derived from the propositions Section 2.2. Table 3 shows examples of this process. Note that these correspond to the text explanations in e-sNLI and e-SNLI-VE datasets. A and B refer to the objects being compared in the hypothesis and premise (only two objects were being compared at each time).

Original $f(u, \text{There are people playing the piano}) \rightarrow \texttt{Contradict}$
Counterfactual $f(u, \text{There are people playing woodwind instruments.}) \rightarrow \texttt{Entail}$
Explanation *"The people in the photo are not playing the piano. They are instead playing other woodwind instruments."*
:= ¬(playing the piano ⇔ playing other woodwind instruments)

Original $f(u, \text{A man in safety equipment is working on a piece of metal.}) \rightarrow \texttt{Entail}$
Counterfactual $f(u, \text{A man in protective gear is working on a piece of metal.}) \rightarrow \texttt{Entail}$
Explanation *"Protective gear is a piece of safety equipment"*
:= (safety equipment ⇔ protective gear)

Original $f(u, \text{Two kids walk home from school.} \rightarrow \texttt{Neutral}$
Counterfactual $f(u, \text{Two kids are walking down a sidewalk.}) \rightarrow \texttt{Entail}$
Counterfactual $f(u, \text{Two kids are going home from school.}) \rightarrow \texttt{Neutral}$
Explanation *"Kids walking down a sidewalk are not necessarily going home from school."*
:= ¬(walking down a sidewalk ⇒ going home from school)

Table 3: Examples of counterfactual hypothesis rewrites from the explanation in SNLI-VE dataset (Do et al. 2020). $f(u, x) \rightarrow y$ shows the expected model prediction given the picture premise (u) and either original or counterfactual hypothesis. Note that `Neutral` can still result in `Neutral` (Section 2.2: Proposition 3).

Axiom 1. Substitution for formulas (Fitting 2012) For any variables A and B and any formula $x_{[A]}$ containing A, if $x_{[B\setminus A]}$ is obtained by replacing any number of free occurrences of A in x with B, then for $A \Leftrightarrow B$, $x_{[A]} \Leftrightarrow x_{[B\setminus A]}$.

Assumption 1. The hypothesis x and premise u contain n free variables, of which the variables A and B are members in u and x. We denote the membership of $A \subseteq x$ as $x_{[A]}$.

Assumption 2. Given $R(A, B)$, the counterfactual hypothesis can be constructed by applying Axiom 1 replacing A with B, denoted $x_{[B\setminus A]} = x^{cf}$.[8]

Assumption 3. The predicate relation expressed in $R(A, B)$ is a sufficient condition, to explain f predicted label.

We formally derive the associated counterfactual label y^{cf} (expected satisfiability result) of the new counterfactual hypothesis x^{cf}. Figure 1 and Table 3 show examples for this process. The proofs follow the following high-level structure:
1. Substitution of variables A and B (Axiom 1 and Assumption 1) to construct x^{cf} (Assumption 2).
2. Examine the predicate relationship $R(A, B)$ and derive the logical relationship between the x^{cf} and the premise u.

[8]Equivalence formulas may be substituted for one another without changing that formula's truth value (Fitting 2012).

Proposition 1. *If the original label is* `E`*, then the associated counterfactual label is* `E`*.*

Proof. By Assumption 1, the logical proposition represented by `E` is $u \Rightarrow x_{[A]}$. Since $A \Leftrightarrow B$, then $u \Rightarrow x_{[B\setminus A]}$, by Axiom 1, $u \Rightarrow x^{\text{cf}}$. Therefore we have the resulting propositional formula and associated counterfactual label $(u \Rightarrow x) \Leftrightarrow (u \Rightarrow x^{\text{cf}})$, i.e. $(y = \texttt{E}) \Leftrightarrow (y^{\text{cf}} = \texttt{E})$. □

Proposition 2. *If the original label is* `C`*, then the associated counterfactual label is* `E`*.*

Proof. By Assumption 1, the logical proposition represented by `C` is $(u \Rightarrow \neg x_{[A]})$. Since $\neg(A \Leftrightarrow B)$ is equivalent to $(A \Leftrightarrow \neg B)$, then by Axiom 1, $u \Rightarrow \neg(x_{[\neg B\setminus A]})$. However it is not possible to test for the "negation of" variables as negation is not seen in training data for NLI. Un-

der Assumption 3,[9] if the explanation $\neg(A \Leftrightarrow B)$ sufficiently explains the label, then having $x_{[B\setminus A]}$ negates (or 'flips') the label to Entailment $u \Rightarrow \neg\neg(x_{[B\setminus A]})$. Therefore we have the result $(u \Rightarrow \neg x) \Leftrightarrow (u \Rightarrow \neg\neg(x^{\text{cf}}))$, i.e., $(y = \texttt{C}) \Leftrightarrow (y^{\text{cf}} = \texttt{E})$. □

Proposition 3. *If the original label is* N*, then there are* ***two*** *associated counterfactual labels,* E*, and* N*.*

Two conditions arise because $\neg(A \Rightarrow B) \Leftrightarrow A$ AND $\neg B$. In the interest of space, we present the proof for Proposition 3 in the supplementary material.[10] Table 2 summarises the original propositional formula in the hypothesis x and premise u, the predicate relations $R(A, B)$ and the associated y^{cf}.

2.3 Proposed Metric: FTC

We introduce the metric Faithfulness-Through-Counterfactuals (FTC), to capture the difference in model predicted probabilities $p(\hat{y}^{\text{cf}})$ from the associated counterfactual label y^{cf}.

$$\text{FTC} = 1 - d(p(\hat{y}^{\text{cf}}), p(y^{\text{cf}})) \quad (1)$$

Choice of Distance Function (d) We consider three metrics for d, $\mathbb{1}[\text{argmax}(p(\hat{y}^{\text{cf}}))! = y^{\text{cf}}]$ denoted FTC-δ, KL Divergence (FTC-$\mathcal{K}$) and Wasserstein distance (FTC-$\mathcal{W}$) with symmetrical distance of 1 between E and C and $0 \geq \alpha \geq 1$ distance between N and E, C.

2.4 Generating Counterfactual Hypotheses

In the previous section, we had assumed that the logical variables A and B are substitutable in x directly. Indeed generating a counterfactual hypothesis would be trivial if A and B could be directly extracted from the explanation, and directly substituted in the hypothesis.[11] However, open domain semantic parsing is an unsolved problem (Lee, Gottschlich, and Roth 2021) of which to our knowledge, there is no off-the-shelf solution which does not require fine-tuning on a train set.Hence, we propose to leverage on advances in few-shot priming (Brown et al. 2020) which only requires several handwritten examples and no further fine-tuning. This is compared against the baseline of parsing via 'extract and transform' using regex (experiment described in Section 3.2).

Extract and Transform with Regex We adopt and extend templates identified by Camburu et al. (2020) for explanations, who noted that these templates are a "natural consequence of the task and dataset". Regex methods perform a rule-based span extraction and replacement (we allow for stemmed word replacement). Extraction rules are shown in supplementary material.

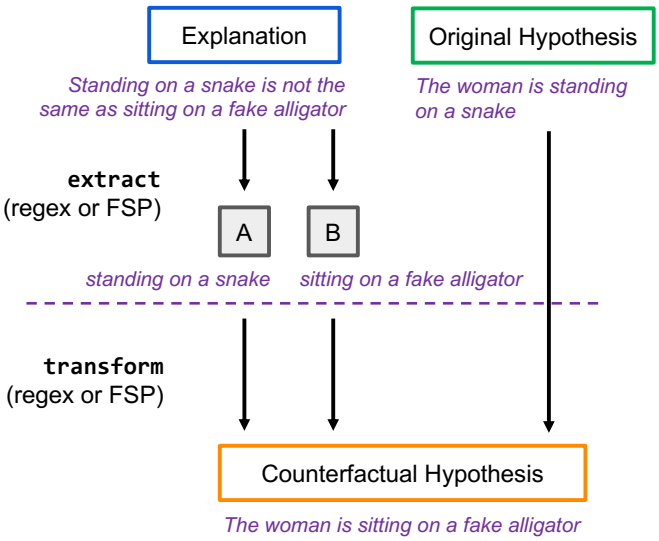

Figure 2: Two step counterfactual rewriting of the hypothesis, according to the explanation. We implement extract and transform steps using regular expressions (regex) and few-shot priming (FSP) models.

Extract and Transform with Few-shot Priming We compare the brittle regex pipeline with the modern paradigm of in-context few-shot priming. This is an attractive option in our setting where we do not have prior training data for generating counterfactual hypotheses, and the task appears to be related to manipulation of text strings.

We thus consider a two step process of 1) extracting the logical spans, i.e., A and B in Table 2 from the explanation, and 2) modifying the hypothesis given these extracted spans.[12] Given a sequence of priming examples of how to extract spans from the explanation in the prefix, the model should perform the extraction given a test explanation.

Reducing Natural Language Artifacts with x^{cf} Previous attempts to test the explanation directly as input to a trained model, are subject to confounds of "label leakage" because of the close association between label and syntatic form of the explanation (Pruthi et al. 2020). Crucially, our proposed method sidesteps this confound by applying *logical satisfiability* checks via predicate relations $R(A, B)$. In theory, the construction of x^{cf} should preserve the syntatic structure of the original hypothesis, while only changing the semantics of A and B.

3 Experiments and Results

We validate the method through three sets of experiments.

i) Evaluating the quality of generated counterfactual hypotheses from a few-shot generator $\mathcal{H}^{\text{model}}$ (Section 3.2).

[9] The violation of Assumption 3, can result in partially 'unfaithful' explanations that do not provide enough information to explain model prediction.

[10] Due to AAAI page limits, technical appendices/supplementary material are available at https://arxiv.org/pdf/2205.12469.pdf and code is available on request.

[11] Consider the hypothesis: "the boy is outside" and the explanation: "A tire swing is usually installed outside". Naive substitution would result in "the boy is a tire swing".

[12] This two step-process is necessary as preliminary experiments show that even few-shot priming models with 175B parameters (academic access GPT3) are not able to construct counterfactual hypothesis in a single step.

ii) Evaluating the proposed metric (FTC) on gold-model agreement of human generated x^{cf}, and comparing this to existing faithfulness metrics in the literature (Section 3.3).

iii) Studying the sensitivity of our proposed approach compared to other metrics given pathological explanations (Section 3.4).

3.1 Experimental Setup

Datasets We consider logical entailment datasets, e-SNLI (Camburu et al. 2018) and e-SNLI-VE (Do et al. 2020) which are the only explainable logical entailment datasets available at point of writing (Wiegreffe and Marasović 2021). e-SNLI consists of crowdsourced explanations for SNLI. e-SNLI-VE replaces the textual premise u of SNLI with Flickr30k images (Young et al. 2014). To avoid trivial word overlap between u and x^{cf}, we adopt the image representation for the premise u in our experiments for Section 3.3 and Section 3.4. Note that Section 3.2 only requires x and z.

Models For the task model f, we adopt a state-of-art multimodal model, `CLIP` (Radford et al. 2021), and fine-tune a 2-layer MLP to train a predictor $f(u, x) \rightarrow y$. For the explanation generator, g, we follow Do et al. (2020) and fine-tune a modified GPT2 (Radford et al. 2019). Training details are in supplementary material. For the counterfactual hypothesis generator, $\mathcal{H}^{\text{model}}$ we adopt a pretrained GPT2-XL and GPT-Neo1.3B and 2.7B (Black et al. 2021) without further fine-tuning, and apply only handwritten prompts. Prompt examples were randomly sampled from the training set and we used 20 prompts for each label.[13]

3.2 Quality of Counterfactual Hypotheses

As the feasibility of our automatic approach depends on the quality of counterfactual generation, we evaluate $\hat{x}^{\text{cf}} \leftarrow \mathcal{H}^{\text{model}}(x, z)$ against gold counterfactuals, $x^{\text{cf}^*} \leftarrow \mathcal{H}^{\text{human}}(x, z)$, where $\mathcal{H}^{\text{human}}$ refers to the human annotator, $\mathcal{H}^{\text{model}}$ refers to our automated hypothesis generator, z refers to explanations and x refers to the original hypothesis. We randomly sample 300 examples from the validation set (100 each for `E`, `C`, `N`) and ask annotators to write counterfactual hypothesis for human generated explanations, z^* and model generated explanations $\hat{z}$. Annotators are asked to revise x such that the logic in z is expressed in the new counterfactual x^{cf} supplementary material. We show annotators the *same* set of examples that were used to prompt $\mathcal{H}^{\text{model}}$. We obtain three annotations per datapoint for multiple reference sentences.[14]

Experiment Conditions As described in Section 2.4, we adopt a two-step process of 'extract' and 'transform' for generating $\hat{x}^{\text{cf}} \leftarrow \mathcal{H}^{\text{model}}(x, z)$. We compare different combinations of either extracting and transforming with regex or $\mathcal{H}^{\text{model}}$. Our main results for the largest and best performing model (GPT-Neo2.7B) is shown in Table 4, and additional results for GPT-Neo1.3B and GPT2-XL in supplementary material.

[13]Further details and example of prompt templates are available in supplementary material.

[14]"A young boy wearing a white shirt on a beach" and "A young boy on a beach wearing a white shirt" are both valid.

Metric The evaluation metric used is METEOR (Banerjee and Lavie 2005), that was found to correlate with human judgement (Kayser et al. 2021). METEOR computes harmonic mean of unigram precision and recall and accounts for stemming. As the validity of any text generation metric is debatable (Deng et al. 2021), we further quantify the downstream effects of the automated process through Section 3.3.

Results (Presented in Table 4).

1. The combination of using models for both extract and transform steps (last row) performs best in all cases with human explanations z^*, and close to best with model explanations, $\hat{z}$.

2. The performance of most methods are better on $\hat{z}$ than z^* which might be explained by the more 'standardised' text format in $\hat{z}$. Brahman et al. (2020) reported that generator models tend to follow a similar format, supporting this interpretation. The row regex-regex can be seen as a direct comparison of how 'standardised' $\hat{z}$ is compared to z^* as it indicates the performance for a brittle rule-based approach.

3. For ease of rewriting each class (column-wise), `C` $>$ `E` $>$ `N` for z^* in most cases, which highlight the relative complexity (`N` tends to be expressed in a less 'straightforward' manner) of extracting and transforming hypothesis with different types of explanations.

3.3 Metric Validation for Predictions on Counterfactual Hypotheses

In this experiment, we compare faithfulness scores with human predictions on counterfactual hypotheses, assuming that humans are consistent with the logically derived answer (we describe the case when they are not in the next section).

As described in Section 1, faithfulness metrics should focus on utility of explanations to the user (Jacovi and Goldberg 2020). One practical utility is that humans should have a better handle of the model's predictions on slightly different input.[15]The domain of new input that we test, is the counterfactual hypotheses $x^{\text{cf}} \leftarrow \mathcal{H}^{\text{human}}(x, z^*)$ that has been generated by humans using human explanations in Section 3.2. Faithfulness metrics should therefore be good at discriminating whether the gold predictions agree with the model predictions.

Since faithfulness metrics are real-valued, and agreement is discrete binary, we are interested in how well they can discriminate between the two groups. We use the Wilcoxon rank-sum test which is a nonparametric test for the null hypothesis that two groups (agreement vs disagreement) are equal. Recall that our FTC metric proposes to use $\mathcal{H}^{\text{model}}$ to generate a counterfactual hypothesis from the explanations. If $\mathcal{H}^{\text{model}}$ had produced exactly the same x^{cf} as the human $\mathcal{H}^{\text{human}}$, then FTC should be very effective at separating the two outcome groups, and the test statistic $\rho \in [0, 1]$ would be very close to 1. However we know this automated process is imperfect (Section 3.2), and the implications of a 0.7 or 0.8 METEOR score are unclear.

[15]Simulatability studies (Doshi-Velez and Kim 2017) using raw explanations were considered, however this is subject to confounds in Label leakage (Pruthi et al. 2020; Hase et al. 2020).

Extract	Transform	Human Explanations $z^* \mapsto x^{\text{cf}}$ C	E	$\text{N}_{[A]}$	$\text{N}_{[B]}$	Generated Explanations $\hat{z} \mapsto x^{\text{cf}}$ C	E	$\text{N}_{[A]}$	$\text{N}_{[B]}$
regex	regex	0.690	0.638	0.010	0.012	0.744	0.701	0.473	0.418
regex	Neo2.7B	0.690	0.709	0.710	**0.752**	0.840	0.860	**0.805**	**0.714**
Neo2.7B	regex	0.801	0.691	0.509	0.565	**0.893**	0.781	0.669	0.584
Neo2.7B	Neo2.7B	**0.822**	**0.782**	**0.743**	**0.754**	0.870	**0.881**	**0.807**	0.709

Table 4: METEOR scores for hypothesis generated by regex, GPT-Neo, or their combination, vs human generated hypothesis. Bold is applied column-wise. We show the breakdown of revised hypothesis by class label (Contradiction C, Entailment E and Neutral N), and also whether the explanations are human or model generated. Note that there are two counterfactual hypothesis generated for N, $\text{N}_{[A]}$ and $\text{N}_{[B]}$ which correspond to $x^{\text{cf}}_{[A]}$ and $x^{\text{cf}}_{[B]}$ as described in Section 2.2: Proposition 3.

	C	E	$\text{N}_{[A]}$	$\text{N}_{[B]}$
LAS	0.585	0.481	0.436	0.410
LRA	0.547	0.645*	0.400	0.572
FTC-δ	0.627	0.634*	0.581	0.655*
FTC-$\mathcal{K}$	0.743*	0.747*	0.631	0.644*
FTC-$\mathcal{W}$	0.692	0.748*	0.605	0.644*

Table 5: ρ-statistic $\in [0, 1]$ for Wilcoxon rank-sum test on different faithfulness metrics. A larger ρ-statistic indicates a larger effect size. LAS: Label Adjusted Simulation, LRA: Label Rationale Agreement, and FTC: Faithfulness-by-Counterfactual (ours), and FTC-$\mathcal{W}$ and FTC-$\mathcal{K}$ ($\alpha = 0.7$) are the Wasserstein and Kl-divergence variants. * indicates significance (p-value< 0.05). The test is conducted between the two groups; data points where the human agrees vs disagrees with the model's prediction on counterfactual hypothesis.

We collect annotations for 100 data instances grouped by each original label-class, and obtain 3 annotations per instance.[16] Annotators are required to rate whether (x^{cf}, u) entails, contradicts, or is neutral (supplementary material).

Results (Presented in Table 5)

1. FTC variants have the highest $\rho \in [0, 1]$ statistic (higher is better and indicates a larger effect size), which indicates that it is the most discriminative metric for whether the human can simulate the model's prediction given new counterfactual inputs. This is expected as the simulation procedure is similar to how FTC is calculated.

2. As reported in Section 3.2 discrepancies in the automated rewriting process affect the scoring of the metric in ways which are not easily captured by sentence generation metrics (Table 4). The results show that the best-automatic rewriting process of adopting GPT-Neo2.7B achieves 0.58 to 0.75 ρ-statistic across C, E, N, giving us an indication of the downstream impact of 0.74 to 0.82 METEOR score.

3. KL-Divergence (FTC-$\mathcal{K}$) and Wasserstein Distance (FTC-$\mathcal{W}$) perform slightly better than the naive Identity function (FTC-δ). This indicates that a soft computation over probabilities can account for some of the errors in the x^{cf} rewriting process.

4. With the exception of LAS on C and LRA on E, we find that other metrics have poorer ρ across C, E, N. This indicates that the metric is relatively poorer at distinguishing between the two outcome groups.

Examining Inconsistent Explanations In cases where the human NLI label does *not* correspond to the logically derived NLI class in Table 2, our method suggests that the human provided explanations themselves are *not* faithful to the human's 'internal NLI model'. supplementary materialshows examples of these cases, which we typically find to be due to explanations of low quality, supporting the central thesis of the paper. In the previous results, we filter out data points where majority of the annotators do not predict the logically derived NLI class. We report the non-filtered results in the supplementary material which demonstrate negative results and a failure case when the 'gold' human explanations are themselves problematic.[17]

3.4 Sensitivity Analysis

As a sanity check, good faithfulness metrics should be sensitive towards unfaithful explanations, i.e. they should perform worse on unfaithful explanations compared to faithful ones. We perform a sensitivity analysis on various metrics by examining their raw scores on unfaithful explanations *by construction*. These are constructed by leaving out all but one type of input to the explanation generator. The 'complete' set of inputs are the entailment label (y), hypothesis (x), and premise (u). Note that the 'complete' set of inputs to the explanation generator does not guarantee faithful explanations, but they are guaranteed to be less pathological than leaving out all but one type of input. We additionally consider BertScore (Zhang et al. 2019) and METEOR (Banerjee and Lavie 2005) which are text similarity metrics evaluated against the human explanation, and use FTC-$\mathcal{K}$ variant which has an upperbound of 1 and a high ρ-statistic in the previous

[16] The inter-annotator agreement, as measured by Fleiss' kappa, equals 0.51. We aggregate the final label using the majority vote.

[17] Quantifying the extent and range of annotation errors for explanations is outside of the scope of this work and we refer readers to Valentino, Pratt-Hartmann, and Freitas (2021) who report that explanations in SNLI are valid logical relations 60% of the time and other times they are problematic due to annotation errors.

x	u	y	BER	MET	LAS	LRA	FTC-$\mathcal{K}$
✓	✓	✓	0.888	0.245	0.047	0.788	0.126
✓			0.885	0.233	-0.035	0.561	-0.200
	✓		0.869	0.120	-0.063	0.298	-0.055
		✓	0.865	0.097	0.068	0.630	-0.245

Table 6: Raw scores for different metrics, by perturbing inputs to the explanation generator (sensitivity analysis). x, u, and y refers to hypothesis, premise, and label respectively. BER: BertScore, MET: METEOR, LAS: Label Adjusted Simulation, LRA: Label Rationale Association, FTC-$\mathcal{K}$ (ours). Note that the first two are measures of text similarity, while the other three are designed to measure the faithfulness.

experiment.[18]

Results (presented in Table 6)

1. FTC-$\mathcal{K}$ performs consistently better for the non-pathological (first row 0.126) vs pathological explanations (-0.200, -0.055, -0.245). The same 'correct' trend is observed for LRA 0.788 vs (0.561, 0.298, 0.630).

2. We find that an off-the-shelf BertScore (Zhang et al. 2019) has a suprisingly low range of values for the different conditions (0.865 to 0.888). METEOR scores conditioned only on x are also very close to the full range of inputs (0.233 vs 0.245) indicating superficial word similarity of the explanations when conditioned on just x.

3. LAS scores are in the 'wrong' direction, namely that explanations generated with all of the relevant inputs perform worse than just having the label.

4 Related Work

We outline existing methods which evaluate free-text generated Natural Language explanations, their assumptions of faithfulness, and describe how they operationalise these assumptions. We focus on LAS and LRA, which are used in our experiments, and provided more discussion of related work in supplementary material.

4.1 Leakage Adjusted Simulation

This method assumes explanations are faithful if they allow a model to be more simulatable. A model is simulatable to the extent that an observer, or simulator, can predict its outputs (Doshi-Velez and Kim 2017; Hase et al. 2020; Kumar and Talukdar 2020). From this perspective, one might use a simulator's (either a human or model) accuracy with explanations as input, to measure explanation quality. However, as Hase et al. (2020) argues, the simulator's success does not reflect explanation quality when the explanation leaks the label to the simulator. They thus propose Leakage-adjusted Simulatability (LAS) which performs a macro-average of leakable and non-leakable explanations. However, a high occurrence of label leakage may overwrite the effect of macro-averaging (Pruthi et al. 2020).

[18]As described in Section 2.3, FTC-$\mathcal{K}$ is $1 -$ the KL term, which has lower bound 0 and no upper bound. Hence the upper bound on FTC-$\mathcal{K}$ is 1.

4.2 Label Rationale Association

According to Wiegreffe, Marasović, and Smith (2021), "at a minimum, rationales must be implicitly or explicitly tied to the model's prediction." Their method tests whether label and explanations are similarly robust to noise in the input. Although designed to be highly generalisable to generative explanations, this assumption may be overly general for more rigorous notions of faithfulness. Consider the scenario where merely changing the label to the generator results in "sufficiently different" explanations being generated (Kumar and Talukdar 2020), whether or not the original explanation was actually faithful, LRA will assign this a high score.

4.3 Counterfactuals as Explanations

Our approach differs from the literature on Counterfactuals as explanations (Mothilal, Sharma, and Tan 2020; Verma, Dickerson, and Hines 2020) as we do not generate counterfactual explanations, but generate counterfactual hypotheses based on the explantions. Camburu et al. (2020) work has a similar flavor, where they "reverse" a hypothesis. However, they focus on show the pathologies of a generator by searching for (adversarial) input hypothesis that cause the model to generate logically inconsistent explanations. Ge et al. (2021) also constructs 'counterfactual inputs', but search for existing features in the original input (applicable only to extractive explanations), while Varshney et al. (2022) also generate hypothesis, but primarily for augmenting training data with sentence transformations.

4.4 Natural Logic vs Free Logic

Previous work on "Natural Logic" (MacCartney and Manning 2007) relies on natural language features to guide inferences. For instance, changing specific terms to more general ones preserves entailment. This sidesteps the difficulties of translating sentences into First-order-Logic. Natural logic systems (Angeli and Manning 2014) have been used in explainable fact verification (Krishna, Riedel, and Vlachos 2021) which constructs "explanations" by presenting logical steps for inference. However these approaches still require a knowledge base to train or mine truth values, e.g, "in Paris" $\subseteq$ "in France". In contrast, our method does not require additional training and is a procedurally lightweight method relying on off-the-shelf pretrained models.

5 Conclusion

Measuring faithfulness of free-text explanations with respect to a task model is a challenging problem due to confounds introduced by testing explanations directly. In this work, we propose an approach to evaluating explanations for NLI tasks which uses the predicate logic expressed in explanations to construct counterfactual hypotheses, and tests the *satisfiability* of the resulting hypothesis. Our experiments on validating counterfactual hypothesis generation and the fully automated approach distinguishing when humans agree and disagree with the model's prediction on counterfactual hypotheses show that a proposed automatic pipeline is a viable approximation to the theoretical method. Further, we show that our metric is sensitive to pathologically unfaithful explanations.

Acknowledgments

We thank Benjamin Van Durme for computational resources for experiments that were ran outside Meta AI. Alexandra Delucia, Marc Marone, Patrick Xia, Matthew Francis-Landau, Lin Chu-Cheng for comments and discussion.

References

Angeli, G.; and Manning, C. D. 2014. Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 534–545.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bencivenga, E. 2002. Free logics. In *Handbook of philosophical logic*, 147–196. Springer.

Black, S.; Gao, L.; Wang, P.; Leahy, C.; and Biderman, S. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. `https://github.com/EleutherAI/gpt-neo`. Accessed: 2021-07-30.

Boolos, G. S.; Burgess, J. P.; and Jeffrey, R. C. 2002. *Computability and logic*. Cambridge university press.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. Lisbon, Portugal.

Brahman, F.; Shwartz, V.; Rudinger, R.; and Choi, Y. 2020. Learning to rationalize for nonmonotonic reasoning with distant supervision. *arXiv preprint arXiv:2012.08012*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Camburu, O.-M.; Shillingford, B.; Minervini, P.; Lukasiewicz, T.; and Blunsom, P. 2020. Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4157–4165. Online.

Danilevsky, M.; Qian, K.; Aharonov, R.; Katsis, Y.; Kawas, B.; and Sen, P. 2020. A Survey of the State of Explainable AI for Natural Language Processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 447–459. Suzhou, China.

Deng, M.; Tan, B.; Liu, Z.; Xing, E.; and Hu, Z. 2021. Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7580–7605. Online and Punta Cana, Dominican Republic.

DeYoung, J.; Jain, S.; Rajani, N. F.; Lehman, E.; Xiong, C.; Socher, R.; and Wallace, B. C. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–4458. Online.

Do, V.; Camburu, O.-M.; Akata, Z.; and Lukasiewicz, T. 2020. e-SNLI-VE-2.0: Corrected Visual-Textual Entailment with Natural Language Explanations. *arXiv preprint arXiv:2004.03744*.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Fitting, M. 2012. *First-order logic and automated theorem proving*. Springer Science & Business Media.

Ge, Y.; Liu, S.; Li, Z.; Xu, S.; Geng, S.; Li, Y.; Tan, J.; Sun, F.; and Zhang, Y. 2021. Counterfactual Evaluation for Explainable AI. *arXiv preprint arXiv:2109.01962*.

Hase, P.; Zhang, S.; Xie, H.; and Bansal, M. 2020. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4351–4367. Online.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. Online.

Kayser, M.; Camburu, O.-M.; Salewski, L.; Emde, C.; Do, V.; Akata, Z.; and Lukasiewicz, T. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1244–1254.

Krishna, A.; Riedel, S.; and Vlachos, A. 2021. ProoFVer: Natural Logic Theorem Proving for Fact Verification. *CoRR*, abs/2108.11357.

Kumar, S.; and Talukdar, P. 2020. NILE : Natural Language Inference with Faithful Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8730–8742. Online.

Lambert, K. 1967. Free logic and the concept of existence. *Notre Dame Journal of Formal Logic*, 8(1-2): 133–144.

Lee, C.; Gottschlich, J.; and Roth, D. 2021. Toward Code Generation: A Survey and Lessons from Semantic Parsing. *CoRR*, abs/2105.03317.

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

MacCartney, B.; and Manning, C. D. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 193–200.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual

explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617.

Nolt, J. 2021. Free Logic. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688.

Pruthi, D.; Dhingra, B.; Soares, L. B.; Collins, M.; Lipton, Z. C.; Neubig, G.; and Cohen, W. W. 2020. Evaluating Explanations: How much do explanations from the teacher aid students? *CoRR*, abs/2012.00893.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4932–4942. Florence, Italy.

Valentino, M.; Pratt-Hartmann, I.; and Freitas, A. 2021. Do Natural Language Explanations Represent Valid Logical Arguments? Verifying Entailment in Explainable NLI Gold Standards. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, 76–86. Groningen, The Netherlands (online).

Varshney, N.; Banerjee, P.; Gokhale, T.; and Baral, C. 2022. Unsupervised Natural Language Inference Using PHL Triplet Generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2003–2016. Dublin, Ireland: Association for Computational Linguistics.

Verma, S.; Dickerson, J.; and Hines, K. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

Wiegreffe, S.; and Marasović, A. 2021. Teach me to explain: A review of datasets for explainable nlp. *arXiv preprint arXiv:2102.12060*.

Wiegreffe, S.; Marasović, A.; and Smith, N. A. 2021. Measuring Association Between Labels and Free-Text Rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10266–10284. Online and Punta Cana, Dominican Republic.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.