Analyzing Semantic Faithfulness of Language Models via Input Intervention on Question Answering

Akshay Chaturvedi
IRIT, Université Paul Sabatier
Toulouse, France
akshay91.isi@gmail.com

Swarnadeep Bhar
IRIT, Université Paul Sabatier
Toulouse, France
swarnadeep.bhar@irit.fr

Soumadeep Saha
Indian Statistical Institute
Kolkata, India
soumadeep.saha97@gmail.com

Utpal Garain
Indian Statistical Institute
Kolkata, India
utpal@isical.ac.in

Nicholas Asher
IRIT, Université Paul Sabatier
Toulouse, France
nicholas.asher@irit.fr

Transformer-based language models have been shown to be highly effective for several NLP tasks. In this article, we consider three transformer models, BERT, RoBERTa, and XLNet, in both small and large versions, and investigate how faithful their representations are with respect to the semantic content of texts. We formalize a notion of semantic faithfulness, in which the semantic content of a text should causally figure in a model's inferences in question answering. We then test this notion by observing a model's behavior on answering questions about a story after performing two novel semantic interventions—deletion intervention and negation intervention. While transformer models achieve high performance on standard question answering tasks,

Action Editor: Mohit Bansal. Submission received: 16 January 2023; revised version received: 11 July 2023; accepted for publication: 17 July 2023.

https://doi.org/10.1162/coli_a_00493

we show that they fail to be semantically faithful once we perform these interventions for a significant number of cases (∼ 50% for deletion intervention, and ∼ 20% drop in accuracy for negation intervention). We then propose an intervention-based training regime that can mitigate the undesirable effects for deletion intervention by a significant margin (from ∼ 50% to ∼ 6%). We analyze the inner-workings of the models to better understand the effectiveness of intervention-based training for deletion intervention. But we show that this training does not attenuate other aspects of semantic unfaithfulness such as the models' inability to deal with negation intervention or to capture the predicate–argument structure of texts. We also test InstructGPT, via prompting, for its ability to handle the two interventions and to capture predicate–argument structure. While InstructGPT models do achieve very high performance on predicate–argument structure task, they fail to respond adequately to our deletion and negation interventions.

1. Introduction

Transformer-based language models such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019b), and so forth, have revolutionized natural language processing (NLP) research, generating contextualized representations that provide state-of-the-art performance for various tasks like part of speech (POS) tagging, semantic role labeling, and so on. The transfer learning ability of these models has discarded the need for designing task-specific NLP systems. The latest incarnation of language models have now excited both the imagination and the fears of researchers (Black et al. 2022; Castelvecchi 2022) and journalists in the popular press; the models and chatbots based on them seem to be able to do code, argue, and tell stories, but they also have trouble distinguishing fact from fiction.[1]

Given their successes and their hold on the public imagination, researchers are increasingly interested in understanding the *inner workings* of these models (Liu et al. 2019a; Tenney et al. 2019; Talmor et al. 2020). In this article, we look at how a fundamental property of linguistic meaning we call *semantic faithfulness* is encoded in the contextualized representations of transformer-based language models and how that information is used in inferences by the models when answering questions. A semantically faithful model will accurately track the semantic content of questions and texts on which the answers to those questions are based. It is a crucial property for a model to have, if it is to distinguish facts about what is expressed in a text or conversation from fiction or hallucination. We will show that current, popular transformer models are not semantically faithful.

This lack of semantic faithfulness highlights potential problems with popular language models trained with transformer architectures. If these models are not semantically faithful, then they will fail to capture the actual semantic content of texts. Operations that we develop in the body of the paper can be used to dramatically alter text content that these language models would not find, leading to errors with potentially important, negative socioeconomic consequences. Even more worrisome is the instability that we have observed in these models and their occasional failure to keep predicate–argument structure straight; if these models cannot reliably return

1 Here is a sample of stories from the *New York Times*: 'The New Chatbots Could Change the World. Can You Trust Them?' (*NYT* Dec. 10, 2022; 'Meet GPT-3. It Has Learned to Code (and Blog and Argue)', *NYT*, Nov 24, 2020; 'The brilliance and the weirdness of ChatGPT', *NYT*, Dec. 5, 2022).

information semantically entailed by textual content, then we can't rely on their predictions in many sensitive areas. Yet such systems are being deployed rapidly in these areas.

In the next section, we discuss the virtues of semantic faithfulness and preview results of experiments that shed light on a model's semantic faithfulness. In Section 3, we discuss the related work. In Section 4, we turn to the dataset and the transformer models that we will use in examining semantic faithfulness. In Sections 5 and 6, we introduce two types of interventions, deletion and negation interventions, that show that the language models lack semantic faithfulness. In Section 5, we also discuss a kind of training that can help models acquire semantic faithfulness at least with respect to deletion intervention. In Section 7, we look at how models deal with predicate–argument structure and with inferences involving semantically equivalent questions. Once again we find that models lack semantic faithfulness. In Sections 5, 6, and 7, we also analyze semantic faithfulness of InstructGPT (Ouyang et al. 2022) via prompting. Finally, we conclude in Sections 8 and 9. The Appendix section (i.e., Section 10) contains several examples of deletion and negation interventions.

2. The Fundamentals: Semantic Faithfulness

The property of interest is **semantic faithfulness**. It relies on a basic theorem of all formal models of meaning in linguistics: The substitution of semantically equivalent expressions within a larger context should make no difference to the meaning of that context.

2.1 The Definition of Semantic Faithfulness

Linguists study meaning in the field of *semantics*, and **formal semantics** is the study of meaning using logical tools. The basic tool of formal semanticists is the notion of a *structure* $\mathfrak{A}$ for a language $\mathcal{L}$ ($\mathcal{L}$ structure) that assigns denotations of the appropriate type to all well-formed expressions of the language (Dowty, Wall, and Peters 1981). A denotation in an $\mathcal{L}$ structure for an $\mathcal{L}$ sentence ϕ is a truth condition, a formal specification of the circumstances in which ϕ is true.

Formal semanticists use such structures to define *logical consequence*, which allows them to study inferences that follow logically from the meaning of expressions. ϕ of $\mathcal{L}$ is a *logical consequence* of a set of $\mathcal{L}$ sentences S iff any $\mathcal{L}$ structure $\mathfrak{A}$ that makes all of S true also makes ϕ true (Chang and Keisler 1990). Formal semanticists equate logical consequence with *semantic entailment*, also known as *semantic consequence*, the notion according to which something follows from a text T in virtue of T's meaning alone.

Let $\Vdash$ denote the intuitive answerhood relation between a question Q and answers ϕ, ψ to Q, where those answers follow from the semantic content of a story or text T or model of its meaning M_T. Let $\models$ denote logical consequence or semantic entailment and let $\leftrightarrow$ denote the material biconditional of propositional logic. A biconditional expresses an equivalence; so $\phi \leftrightarrow \psi$ means that ϕ will be true just in case ψ is also true. For example, "A goes to a party" $\leftrightarrow$ "B goes to a party" holds if and only if either both A and B or neither go to the party.

Definition [Semantic Faithfulness]: If $T \models \phi \leftrightarrow \psi$, and $T \models Q \leftrightarrow Q'$, then M_T is a semantically faithful model of T iff:

$$T, Q \Vdash \phi \text{ iff } M_T, Q \Vdash \psi \tag{1}$$

and

$$M_T, Q \mid\!\models \psi \text{ iff } M_T, Q' \mid\!\models \psi \tag{2}$$

Note that if $T \models Q \leftrightarrow Q'$ and $T \models \phi \leftrightarrow \psi$, then by the substitution of equal semantic values, it follows in formal semantics that $T, Q \mid\!\models \phi$ iff $T, Q' \mid\!\models \psi$. So in particular, this implies that $T, Q \mid\!\models \phi$ iff $M_T, Q \mid\!\models \psi$. In the rest of the article, we will concentrate on the particular case of semantic faithfulness where $\phi = \psi$.

A semantically faithful machine learning (ML) model of meaning and question answering bases its answers to questions about T on the intuitive, semantic content of T and should mirror the inferences based on semantic consequence: If T's semantic content doesn't support an answer ϕ to question Q, then the model shouldn't provide ϕ in response to Q; if T's semantic content supports an answer ϕ to question Q, then the model should provide ϕ in response to Q. Furthermore, if T is altered to T' so that while $T, Q \mid\!\models \psi$, $T', Q \not\mid\!\models \psi$, a semantically faithful model should replicate this pattern: $M_T, Q \mid\!\models \psi$, but $M_{T'}, Q \not\mid\!\models \psi$. Thus, semantic faithfulness is a normative criterion that tells us how machine learning models of meaning should track human linguistic judgments, when textual input is altered in ways that are relevant to semantic meaning and semantic structure.

Linguistic meaning and the semantic consequence relation $\models$ are defined recursively over semantic structure. Thus, semantic faithfulness provides an important window into machine learning models' grasp of semantic structure and its exploitation during inference. In this, semantic faithfulness goes far beyond and generalizes consistency tests based on lexical substitutions (Elazar et al. 2021a). If a model is not semantically faithful, then it doesn't respect semantic entailment. This in turn means that the model is not capturing correctly at least some aspects of semantic structure. Semantic structure includes predicate–argument structure (i.e., which object described in T has which property) but also defines the scope of operators like negation over other components in the structure. Semantic structure is an essential linguistic component for determining semantic faithfulness, since the semantic structure of a text is the crucial ingredient in recursively defining the logical consequence $\models$ relation that underlies semantic faithfulness.

To see how predicate–argument structure links up with semantic faithfulness, suppose T is the sentence "*a blue car is in front of a red house*". If M_T does not respect predicate–argument structure, then given the question "*was the car red?*", the model may reply yes, simply because it has seen the word "red" in the text. Respecting predicate–argument structure would dictate that the model predicts *no* for the question. Semantic structure also links semantic faithfulness with inference, as we exploit that structure to define valid inference. The lack of semantic structure can cause the model to perform invalid inferences.

2.2 A Remark on Language Models and Formal Semantics

Semantic faithfulness makes use of a traditional notion of logical consequence, which itself depends on truth conditional semantics, in which sentential meaning is defined in terms of the conditions under which the sentence is true. So for instance, the meaning of *there is a red car* is defined in terms of the situations in which there is a car that is red (Davidson 1967). All this seems distant from the distributional view of semantics that informs language models (LMs). However, the two are in fact complementary (Asher

2011). Fernando (2004) shows how to provide a semantics of temporal expressions that fully captures their truth conditional meaning using strings of linguistic symbols or continuations. Graf (2019) shows how to use strings to define generalized quantifiers, which include the meanings of determiners like *every, some, finitely many* (Barwise and Cooper 1981). In such examples of continuation semantics, meanings are defined using strings of words or other symbols, similarly to distributional semantics, or functions over such strings. When defined over the appropriate set of strings, continuation semantics subsumes the notion of logical consequence $\models$ under certain mild assumptions (Reynolds 1974).

Inspired by this earlier literature, Asher, Paul, and Venant (2017) provide a model of language in terms of a space of finite and infinite strings. Many of these strings are just non-meaningful sequences of words but the space also includes coherent and consistent strings that form meaningful texts and conversations. Building on formal theories of textual and conversational meaning (Kamp and Reyle 1993; Asher 1993; De Groote 2006; Asher and Pogodalla 2010), Asher, Paul, and Venant (2017) use this subset of coherent and consistent texts and conversations to define the semantics and strategic consequences of conversational moves in terms of possible continuations.

Such continuation semantics also provides a more refined notion of meaning compared to that of truth conditional semantics. Consider, for instance, the set of most probable continuations for Example (1-a). They are not the most probable continuations for (1-b), even though (1-a) and (1-b) have the same truth conditional meaning.

(1) a. If we do this, 2/3 of the population will be saved.

b. If we do this, 1/3 of the population will die.

(1-a)'s most likely continuations focus perhaps on implementation of the proposed action; (1-b)'s most likely continuations would focus on a search for other alternatives. Thus, while (1-a) and (1-b) are semantically equivalent with respect to denotational, truth conditional semantics, they do not generate the same probability distribution over possible continuations and so have arguably distinct meanings in a continuation semantics that takes the future evolution of a conversation or text into account. Thus, continuation semantics is a natural and non-conservative extension of truth conditional semantics.

LMs find their place rather naturally in this framework (Fernando 2022). LMs provide a probability distribution over possible continuations. Thus, they are sensitive to and can predict possible continuations of a given text or discourse. LMs naturally can be seen as providing a continuation semantics. In principle continuation semantics as practiced by LMs can in principle capture a finer grained semantics than denotational truth conditional semantics, as well as pragmatic and strategic elements of language.

For an LM to be semantically faithful and to produce coherent texts and conversations, it must learn the right probability distribution over the right set of strings. That is, it has to distinguish sense from nonsense, and it has then to recognize inconsistent from consistent strings, incoherent from coherent ones. If it does so, then the LM will have mastered both $\models$ and the more difficult to capture notion of semantic coherence that underlies well-formed texts and conversations. If it does not do so, it will not be semantically faithful. The fact that continuation semantics subsumes the logical consequence relation $\models$ of formal and discourse semantics reinforces our contention that semantic faithfulness based on such a semantics should be a necessary constraint

for adequate meanings based on continuations or meanings based on distributions. Semantic faithfulness is not only a test for an adequate notion of meaning but it also offers a road towards training LMs to better reflect an intuitive notion of semantic consequence and coherence.

In designing experiments to test semantic faithfulness and LM model inference then, we need to pay attention to continuation semantics and how possible interventions can affect discourse continuations. Our interventions exploit the semantics of continuations. We need to do this, because LMs are sensitive to continuations and can detect low probablity continuations. Simple insertions of materials to affect semantic content threaten to not end up testing the inferences we want to test but rather signal an LM's sensitivity to low probability continuations. Continuation semantics provides a rationale for human in the loop constructions of interventions that respect or shift semantic content and continuations (Kaushik, Hovy, and Lipton 2020; Gardner et al. 2020).

2.3 A Summary of Our Contributions

We show that transformer representations of meanings are not semantically faithful, and this calls into question their grasp of semantic structure. We detail three types of experiments in which we show large language models fail to be semantically faithful: In the first case, transformers "hallucinate" responses to questions about texts that are not grounded in their semantic content; in the second, models fail to observe modifications of a text that renders it inconsistent with the model's answer to a question about the original text; in the third, we show that models don't reliably capture predicate–argument structure. These are serious problems, as it means that we cannot offer guarantees that these sophisticated text understanding systems capture basic textual, semantic content. Hence, simple semantic inferences cannot be fully trusted. We analyze the reasons for this and suggest some ways of remedying this defect.

To investigate semantic faithfulness of a model M, we look at inferences M must perform to answer a question, given a story or conversation T and a conversation history containing other questions. Table 1 shows an example. We look at question answering before and after performing two new operations, *deletion intervention* and *negation intervention*, that affect the semantic content of T. We look at the particular case of the definition of semantic faithfulness where $\phi = \psi$—that is, we look at whether the same ground-truth answer to a question is given before and after interventions.

Table 1
An example from CoQA dataset (Reddy, Chen, and Manning 2019). XLNet (Yang et al. 2019) correctly predicts *no* for the question *"Did she live alone?"*. However, it still predicts *no* when the rationale (i.e., text marked in **bold**) is removed from the story (i.e., deletion intervention).

Story	Once upon a time, in a barn near a farm house, there lived a little white kitten named Cotton. Cotton lived high up [...] farmer's horses slept. **But Cotton wasn't alone in her little home above the barn, oh no.**
Conversation History	What color was Cotton? white Where did she live? in a barn
Question	Did she live alone?
Prediction	no

Deletion intervention removes from T a text span conveying semantic information necessary and sufficient given T for answering a question with answer ψ. We call the text conveying the targeted semantic information the *rationale.* T itself supports ψ as an answer to Q—in the formalism of equation 1, $T, Q \models \psi$. But post intervention T, call it $d(T)$, does not: $d(T), Q \not\models \psi$. The semantic content of $d(T)$ no longer semantically supports ψ. A semantically faithful model M_T should mirror this shift: $M_T, Q \models \psi$ but $M_{d(T)} \not\models \psi$—which accords with human practice and intuition.

Negation intervention modifies a text T into a text $n(T)$ such that $n(T)$ is inconsistent with ψ, where ψ was an answer to a question supported by the original text. In the formal terms we have used to define semantic faithfulness, $T, Q \models \psi$ but $n(T), Q \models \neg\psi$. One simple instance of negation intervention would insert a negation with scope over the Q targeted semantic information. But this is not the only or even the primary way; in fact such simple cases of negation intervention amount to only 10% of our interventions. To preserve T's discourse coherence and style, changing the content of a text so as to flip the answer in a yes/no question typically requires other changes to T. To consider a simple example, suppose that in Table 1, we consider as our question Q: *was Cotton white?* Performing negation intervention on the rationale, *there lived a little white kitten named Cotton*, led us to replace the rationale with two sentences: *there lived a little kitten named Cotton. Cotton was not white.*

In general, negation intervention tests whether an ML model is sensitive to semantic consistency. A semantically faithful model should no longer answer Q with *yes* post negation intervention. Once again negation intervention exploits the notion of semantic faithfulness. We should observe a shift in the ML model's behavior after negation intervention on a text T, $n(T)$: Supposing that on $T, Q \models \psi$, a semantically faithful model M_T should be such that $M_T, Q \models \psi$ but $M_{n(T)} \not\models \psi$.

Deletion and negation interventions allow us to study the models' behavior in a counterfactual scenario. Such counterfactual scenarios are crucial to understanding the causal efficacy of the rationale in the models' inferring of the ground truth answer for a given question (Schölkopf 2019; Kusner et al. 2017; Asher, Paul, and Russell 2021). Scientific experiments establish or refute causal links between A and B by seeing what happens when A holds and what happens when $\neg A$ holds. Generally, A causes B only if both A and B hold and the counterfactual claim, that if $\neg A$ were true then $\neg B$ would also be true, also holds. So if we can show for intervention i on T that a model M_T is such that $M_T, Q \models \psi$ and $T, Q \models \psi$ but also such that $M_{i(T)}, Q \models \psi$ and $i(T), Q \not\models \psi$—that is, $i(T)$ no longer contains information α (originally in T) that is needed to support ψ as an answer to Q—then we have shown that information α is not causally involved in the inference to ψ.

We perform our experiments on two question answering (QA) datasets, namely, CoQA (Reddy, Chen, and Manning 2019) and HotpotQA (Yang et al. 2018). CoQA is a conversational QA dataset whereas HotpotQA is a single-turn QA dataset. Both the datasets include, for each question, an annotated rationale that human annotators determined to provide the ground truth answers to questions and from which ideally the answer should be computed in a text understanding system. The **bold text** in Table 1 is an example from CoQA of a rationale. We exploited these annotated rationales to study the language models' behavior under our semantic interventions. More precisely, we ask the following question: *Do language models predict the ground truth answer even when the rationale is removed from the story under deletion intervention or negated under negation intervention?*

The surprising answer to our question is "yes". This shows that the models are not semantically faithful. Intuitively, a model should not make such a prediction post

deletion or negation intervention, since the content on which the ground truth answer should be computed (i.e., the rationale) is no longer present in the story. Our interventions show that the rationale is not a cause of the model's computing the ground truth answer; at least they are not necessary for computing the answer. This strongly suggests that such language models are not guaranteed to be semantically faithful, something we establish in greater detail in Sections 5 and 6.

In a third set of experiments in Section 7, we query the models directly for their knowledge of predicate–argument structure in texts. We construct sentences with two objects that each have a different property. We then perform two experiments. In the first, simple experiment, we simply query the model about the properties those objects have. In some cases, some models had trouble even with this simple task. In a second set of experiments, we query the model with two distinct but semantically equivalent yes/no questions. This experiment produces some surprising results where models have trouble answering semantically equivalent questions in the same way, once again indicating a lack of semantic faithfulness. Formally we have:

- two questions, Q, Q',
- $T \models Q \leftrightarrow Q'$ and
- $T, Q \mid\mid\!\!\models \psi$ *iff* $T, Q' \mid\mid\!\!\models \psi$
- but it's **not** the case that $M_T, Q \mid\mid\!\!\models \psi$ *iff* $M_T, Q' \mid\mid\!\!\models \psi$.

Working with *base* and *large* variants of three language models, BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019b), and XLNet (Yang et al. 2019), on the CoQA dataset (Reddy, Chen, and Manning 2019), we make the following five contributions:

1. We show that, despite the models' high performance on CoQA and HotpotQA, they wrongly predict the ground truth answer post deletion intervention for a large number of cases ($\sim$ 50%).
2. We show that a simple intervention-based training strategy is extremely effective in making these models sensitive to deletion intervention without sacrificing high performance on the original dataset.
3. We quantitatively analyze the *inner-workings* of these models by comparing the embeddings of common words under the two training strategies. We find that under intervention-based training, the embeddings are more contextualized with regards to the rationale.
4. For negation intervention, we show that all the models suffer a $\sim$ 20% drop in accuracy when the rationale is negated in the story.
5. We show that, in general, the models have difficulty in capturing predicate–argument structure by examining their behavior on paraphrased questions.
6. We also test the ability of InstructGPT (Ouyang et al. 2022) (i.e., *text-davinci-002* and *text-davinci-003*) to tackle the two interventions and capture predicate–argument structure via prompting. For the two

interventions, InstructGPT models also display similar behavior as the other models. With regards to predicate–argument structure, the models achieve very high performance. However, for certain cases, the models do exhibit inconsistent behavior, as detailed in Section 7.

3. Related Work

There has been a significant amount of research analyzing language models' behavior across different NLP tasks (Rogers, Kovaleva, and Rumshisky 2020). **Probing** has been a popular technique to investigate linguistic structures encoded in the contextualized representations of these models (Pimentel et al. 2020; Hewitt and Liang 2019; Hewitt and Manning 2019; Chi, Hewitt, and Manning 2020). In probing, one trains a model (known as a *probe*) that takes the frozen representations of the language model as input for a particular linguistic task. The high performance of the probe implies that the contextualized representations have encoded the required linguistic information.

In particular, predicate–argument structure has been a subject of probing (Conia and Navigli 2020, 2022). Though most, if not all, of the effort is devoted to finding arguments of verbal predicates denoting actions or events using semantic role labeling formalisms (Chi, Hewitt, and Manning 2020; Conia and Navigli 2020). Little effort has been made in the literature to investigate the grasp of predicate–argument structure at the level of the formal semantic translations of natural language text—which includes the arguments of verbal predicates but also things like adjectival modification.

One major disadvantage of probing methods is that they fail to address how this information is used during inference (Tenney, Das, and Pavlick 2019; Rogers, Kovaleva, and Rumshisky 2020). Probing only shows that there are enough clues in the representation so that a probe model can learn to find, say, the predicate–argument from the language model's representation. It tells us little as to whether the model leverages that implicit information in reasoning about textual content. Our experiments are designed to do the latter.

Another approach to understanding the inner workings of language models studies their behavior at inference time. Elazar et al. (2021b) explores an intervention-based model analysis, called **amnesic probing**. Amnesic probing performs interventions on the hidden representations of the model in order to remove specific morphological information. In principle one could extend this approach to other kinds of linguistic information. Amnesic probing is unlike our work, in which the interventions are performed on the input linguistic content and form. Balasubramanian et al. (2020) showed in related work that BERT is *surprisingly brittle* when one named entity is replaced by another. Sun et al. (2020) showed the lack of robustness of BERT to commonly occurring misspellings.

For the task of question answering, a transformer-based language model with multiple output heads is typically used (Hu et al. 2019). An output head caters to a particular *answer type*. Thus, the usage of multiple output heads allows the model to generate different answer types such as span, yes/no, number, and so forth. Geva et al. (2021) studied the behavior of *non-target* heads, that is, output heads not being used for prediction. They showed that, in some cases, non-target heads are able to explain the models' prediction generated by the *target head*. Schuff, Adel, and Vu (2020) analyzed the QA models that predict answer as well as an explanation. For such models, they manually analyzed the predicted answer and explanation to show that the explanation is often not suitable for the predicted answer. Their methodology is in contrast to our

work, since we simply argue that the model uses the rationale for predicting the answer if it is sensitive to *deletion intervention*.

Researchers in prior work have also studied the behavior of the model on manipulated input texts (Balasubramanian et al. 2020; Sun et al. 2020; Jia and Liang 2017; Song et al. 2021; Belinkov and Bisk 2018; Zhang et al. 2020). However, they usually frame the task in an *adversarial scenario* and rely either on an attack algorithm or complex heuristics for generating manipulated text. The objective in such work is to fool the model with the manipulated text so that the model changes its predictions whereas a human would not change the prediction in the face of the manipulated data.

In contrast, deletion intervention is a simple *content deletion* strategy; it is not designed to get the model to shift its predictions in cases where a human would not. It's not designed to trick or fool ML models. Deletion intervention manipulates the text to test how the deletion of content affects inference; ideally both humans and the ML model should shift their predictions in a similar way given a deletion intervention. Nevertheless, it is also reasonable to expect a model that was successfully attacked in an adversarial setting to be sensitive to deletion intervention.

With respect to negation intervention, researchers have examined the effects of negation and inference at the sentential level on synthetic datasets (Naik et al. 2018; Kassner and Schütze 2020; Hossain et al. 2020; Hosseini et al. 2021). Our aim is more ambitious; we study how transformer models encode both the content C in a text and content C' in a negation-intervened text that is inconsistent with C. Using negation intervention, we test how replacing C with C' affects inference in natural settings. As with deletion intervention, we offer another way of changing the meaning of texts that should make both humans and semantically faithful models change their predictions. There is similar work relevant to negation intervention—on contrast set data and also counterfactual data (Kaushik, Hovy, and Lipton 2020; Gardner et al. 2020). The datasets on which Kaushik, Hovy, and Lipton (2020) and Gardner et al. (2020) operate are less complex discursively than the CoQA dataset. Generally, our tests also go beyond tests of consistency as in Elazar et al. (2021a), which is based on lexical substitutions. Semantic faithfulness, which we test with deletion and negation interventions as well as questions paraphrases, works at a structural level in semantics; it generalizes Elazar et al.'s (2021a) notion of consistency. Also, while Elazar et al. (2021a) focused on pretrained language models, our work is focused on language models finetuned for question answering.

Our study on CoQA and HotpotQA also allows us to look more closely at what the models are actually sensitive to in a longer text or story. We return to this issue in more detail in Section 6. In general, interventions are an important mechanism to build counterfactual models as Kaushik, Hovy, and Lipton (2020) also argue. These are important for understanding causal structure (Schölkopf 2019; Kusner et al. 2017; Barocas, Hardt, and Narayanan 2019).

4. Dataset Specification and Model Architecture

We now describe the two datasets, CoQA and HotpotQA, and the architecture of the models used for this work along with implementation details.

4.1 Datasets

The CoQA dataset consists of a set of stories paired with a sequence of questions based on the story. To answer a particular question, the model has access to the story and previous questions with their ground truth answers—this is the conversation history.

Table 2
Data statistics for CoQA and HotpotQA along with percentage of *unknown* questions.

Dataset	Split	Story	Questions	unk%
CoQA	train	7,199	108,647	1.26
	dev	500	7,983	0.83
HotpotQA	train	84,579	90,447	–
	dev	7,350	7,405	–

The dataset contains questions of five types: *yes/no* questions; questions whose direct answer is a *number*; alternative, or *option* questions (e.g., *do you want tea or coffee?*); questions with an *unknown* answer; and questions whose answer is contained in a *span* of text. The *span* answer type accounts for the majority of the questions ($>$ 75%). The dataset also contains a *human annotated rationale* for each question.

The HotpotQA is a standard (i.e., single-turn) QA dataset. For each question, the dataset contains 2 gold, and 8 distractor Wikipedia paragraphs. Only the gold paragraphs are relevant to the question. The annotated rationale highlights the particular sentences within the gold paragraphs that are needed to answer the question. Given that the language models used in this work have an input limit of 512 tokens, we only feed the two gold paragraphs as input to the model. For the sake of consistency with CoQA, we refer to this concatenated input as **story**. HotpotQA mostly contains span answer type questions ($>$ 90%) and unlike CoQA, this dataset doesn't contain any question with an *unknown* answer. Since the test set for the two datasets is not publicly available, we report the performance of the models across different experimental settings on the development set. Table 2 provides statistics for the training and development set for the two datasets.

4.2 Models

We conducted experiments on *base* and *large* variants of three transformer-based language models—BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019b), and XLNet (Yang et al. 2019). For the sake of consistency, we used the same model architecture for the two datasets. The only difference lies in the input given to the model. For CoQA, to predict the answer for the i^{th} question, Q_i, for a given story S, the models use previous questions along and their ground truth answers from the conversation history. Therefore, the input for the models for the story, S, and question, Q_i, is as follows.

$$\text{XLNet:}[S \texttt{ <sep> } Q_{i-2}A_{i-2}Q_{i-1}A_{i-1}$$
$$Q_i \texttt{ <sep> <cls>}]$$
$$\text{BERT/RoBERTa:}[\texttt{<cls> } Q_{i-2}A_{i-2}Q_{i-1}A_{i-1}Q_i$$
$$\texttt{<sep> } S \texttt{ <sep>}]$$

where <sep> token is used to demarcate the story and the question history, <cls> is a special token which is used for predicting answer for non-span type questions, and A_j denotes the ground truth answer for the question, Q_j. In the rest of the article, we refer to the string $Q_{i-2}A_{i-2}Q_{i-1}A_{i-1}Q_i$ as *question context*. Since HotpotQA is a single-turn

QA dataset, instead of feeding the question content, we only feed the current question Q_i to the model. The rest of the input remains the same.

We used the publicly available XLNet model for this article.[2] The model contains output heads for *unknown*, *yes*, *no*, *number*, *option*, and *span*. Each output head is fed with a concatenation of the CLS embedding and contextualized embeddings of the story weighted by the predicted start probabilities to predict a score.

For BERT and RoBERTa, we implemented the rationale tagging multi-task model described in Ju et al. (2019). Unlike XLNet, the two models are trained on QA as well as on the rationale tagging task. Furthermore, for a question, the two models can predict *yes*, *no*, *unknown*, and *span*. As a result, for CoQA, the two models predict span for 78.9% of the questions in the development set, whereas XLNet predicts span for 75.8%. Similarly, for HotpotQA, the two models predict span for 93.8% of questions in the development set, whereas XLNet predicts span for 91.1%. For span prediction, the start and end logits for the answer are predicted by applying a fully connected layer to the contextualized representation of the story obtained from the last layer of the model.

The rationale tagging task requires predicting whether a token $t \in S$ belongs to the rationale. Let $h_t \in \mathbb{R}^d$ denote the contextualized embedding obtained from the last layer for token t. The model assigns a probability p_t for t to be in the rationale as follows.

$$p_t = \sigma(uReLU(Vh_t)) \tag{3}$$

where $u \in \mathbb{R}^{1\times d}$, $V \in \mathbb{R}^{d\times d}$, $ReLU$ is the rectified linear unit activation function, and $\sigma(.)$ denotes the sigmoid function. An attention mechanism is then used to generate a representation, q^L, as shown below.

$$p'_t = p_t \times h_t \tag{4}$$

$$a_t = softmax(w_1 ReLU(W_2 p'_t)) \tag{5}$$

$$q^L = \sum_t a_t \times p'_t \tag{6}$$

where $w_1 \in \mathbb{R}^{1\times d}$, $W_2 \in \mathbb{R}^{d\times d}$. Let $h_{CLS} \in \mathbb{R}^d$ denote the CLS embedding obtained from the last layer. h_{CLS} is concatenated with the embedding q^L. The concatenated embedding is then used in BERT and RoBERTa to generate a score for *yes*, *no*, and *unknown*, respectively.

4.3 Implementation Details

We implemented the three language models in PyTorch using the HuggingFace library (Wolf et al. 2020). The models were finetuned on the CoQA dataset for 1 epoch. The *base* variant of the three models was trained on a single 11 GB GTX 1080 Ti GPU, whereas the *large* variant was trained on a single 24 GB Quadro RTX 6000 GPU. The code and the additional datasets created as part of this work are publicly available.[3]

2 https://github.com/stevezheng23/mrc_tf.

3 https://github.com/akshay107/sem-faithfulness.

Table 3
An example from CoQA dataset where the rationale (shown in **bold**) is not necessary to answer the question. The question can be answered using the italicized text.

Story	Characters: Sandy, Rose, Jane, Justin, Mrs. Lin [...] Jane: Sandy, I called you yesterday. Your mother told me [...] This year is very important to us. **Sandy:(Crying) My father has lost his job**, and we have no money to pay all the spending. [...] *Jane: Eh...I hear that Sandy's father has lost his job*, and Sandy has a part-time job.
Question	Who was unemployed?
Prediction	Sandy's father

5. Deletion Intervention and Results

In this section, we explain the operation of **deletion intervention** and discuss the proposed intervention-based training. Deletion intervention is an operation that removes the *rationale* of a question *Q* from the *story*. For a few instances in the CoQA dataset, we found that the annotated *rationale* for *Q* was not necessary for answering Q, because the sentences following the rationale contained the relevant information for supplying an answer to *Q*. One such instance is shown in Table 3. In our experiments with CoQA, we did not find any instance where the sentences preceding the rationale contained the necessary information for answering the question. To avoid problems with such examples containing redundancies, given an original story, we created three additional datasets:

1. TS: In this dataset, we truncate the original story (OS) so that the statement containing the rationale is the last statement. We refer to this dataset as TS (short for *truncated story*). The stories in TS do not reduplicate elsewhere information in the rationale. TS has the same number of samples as OS.
2. TS-R: Given TS, we perform *deletion intervention* by removing all the sentences containing the rationale. The cases where the rationale begins from the first sentence itself are discarded. For questions where the model predicts a *span*, we add the ground truth answer (if not already present) post deletion intervention. This is necessary since for the *span* type questions, the model can only predict the ground truth answer if it is present in the story. As an example, consider the question "Where does Alan go after work?" and the story "Alan works in an office. **He goes to a nearby park after work.**" (rationale shown in bold). In this case, TS-R will be "Alan works in an office. park." Since TS-R doesn't contain the information necessary for answering the question, the model should predict *unknown* for such instances. TS-R had a total of 98,436 samples for training and 7,271 samples for evaluation.

3. TS-R+Aug: This dataset is similar to TS-R. However, in this dataset, instead of simply adding the ground truth answer at the end for the aforementioned cases, we use the OpenAI API (i.e., *gpt-3.5-turbo*) to generate a sentence containing the ground truth answer. This generated sentence is then added at the end for the concerned cases. Given the budgetary constraints involved in using the OpenAI API, we only use this dataset for evaluation, and not during training. Generating sentence using the API is not essential for deletion intervention. Rather, the purpose of this dataset is to show that the proposed intervention-based training doesn't solely rely on superficial cues of TS-R to tackle deletion intervention. Out of the 5,351 cases of span questions where ground truth answer was absent post deletion intervention, the API was successfully able to generate sentence containing the ground truth answer for 4,329 cases. Unsuccessful cases were discarded from this dataset. Similar to TS-R, the model should predict *unknown* for TS-R+Aug. This dataset had a total of 6,249 samples for evaluation.

For HotpotQA, given an OS, we constructed two additional datasets: OS-R, OS-R+Aug in a similar fashion. Unlike Table 3, we did not find any such problematic cases for HotpotQA. For OS-R+Aug, out of 5,855 cases of span type questions where the ground truth answer was absent post deletion intervention, *gpt-3.5-turbo* was successful in generating a sentence containing the ground-truth answer for 5,457 cases. Similar to TS-R+Aug, the OS-R+Aug dataset is also used solely for evaluation. OS-R had a total of 87,234 samples for training and 7,119 samples for evaluation. OS-R+Aug had a total of 6,721 samples for evaluation.

We trained the models on the OS dataset and evaluated them on the aforementioned datasets. We refer to this training strategy as OT (short for original training). The OT in Tables 4 and 5 shows EM (exact match), F1, and the percentage of unknown predictions (unk%) of the models for this training strategy on CoQA and HotpotQA, respectively. Note that, for all the datasets, the EM and F1 in the two tables is with respect to the original ground-truth answer (i.e., ground truth answer for OS). As we can see from OT in Table 4, for the datasets OS and TS, the performance of all the models is pretty similar. The performance drops for TS-R, which shows some sensitivity to *deletion intervention*. However, all the models still achieve an EM of $\sim 50\%$, which is intuitively way too high for a semantically faithful model. We believe this shows that the models rely on superficial cues for predicting the answer; for example, in the presence of a question like "What color was X?" it searches for a color word not too far from a mention of X. We also find that, for TS-R, the unk% for RoBERTa, XLNet is significantly higher than BERT. There is a drop in EM and F1 for TS-R+Aug in comparison to TS-R (especially for RoBERTa) but the unk% remains similar, which shows that the model is not adept at handling this dataset. For HotpotQA, the OT in Table 5 highlights the models' inability to tackle deletion intervention even further. We can see that for OS and OS-R, the performance drop is not so apparent. In fact, RoBERTa actually achieves a higher EM on OS-R than OS. Overall, all the models have very high scores on OS-R, which is undesirable. For OS-R+Aug, we see that BERT and XLNet-base have higher scores in comparison to OS-R, whereas the scores drop for the other models like RoBERTa. However, like CoQA, the unk% remains similar.

Apart from the six models, we also evaluated the behavior of InstructGPT (Ouyang et al. 2022) using OpenAI API on deletion intervention. We looked at two InstructGPT

Table 4
EM, F1 score, and percentage of unknown predictions (i.e., *unk*%) of the models under the two training strategies for CoQA.

Model	Dataset	OT			IBT		
		F1	EM	unk%	F1	EM	unk%
BERT-base	OS	76.1	66.3	1.97	76.4	67.2	3.82
	TS	77.2	67.1	2.18	77.7	68.0	7.93
	TS-R	55.6	48.2	1.98	5.7	5.4	93.08
	TS-R+Aug	51.5	44.3	7.10	42.4	36.3	45.50
BERT-large	OS	80.7	71.1	2.01	78.8	69.8	4.20
	TS	81.6	72.1	2.32	80.1	70.7	7.34
	TS-R	63.6	57.8	3.79	5.4	5.1	94.25
	TS-R+Aug	53.4	46.3	8.80	38.3	32.9	51.88
RoBERTa-base	OS	80.3	70.8	1.95	81.2	71.6	2.86
	TS	80.8	71.1	2.64	81.9	72.0	5.20
	TS-R	55.5	51.1	16.92	5.5	5.3	94.25
	TS-R+Aug	39.2	28.2	14.26	6.3	6.0	92.13
RoBERTa-large	OS	87.0	77.7	1.74	86.2	76.9	2.66
	TS	86.8	77.3	2.72	86.3	76.7	4.01
	TS-R	59.9	55.7	22.36	5.1	5.0	95.34
	TS-R+Aug	42.9	32.0	22.18	6.0	5.7	93.65
XLNet-base	OS	82.5	74.8	1.08	81.3	74.2	4.63
	TS	82.1	74.2	1.11	79.6	72.4	10.87
	TS-R	53.5	48.0	14.00	6.6	6.4	93.86
	TS-R+Aug	50.7	45.3	13.45	25.2	22.2	68.75
XLNet-large	OS	86.3	78.9	0.86	83.1	75.8	5.10
	TS	85.6	78.5	2.58	81.0	74.1	10.69
	TS-R	48.1	44.3	31.68	5.6	5.5	95.42
	TS-R+Aug	46.9	42.4	26.18	22.1	19.5	73.93

models: *text-davinci-002* and *text-davinci-003*. As per the models' documentation,[4] the two models are similar. The only difference being that *text-davinci-002* was trained with supervised learning instead of reinforcement learning. For CoQA, we provided the story, the question context (i.e., two previous questions with their ground truth answer), and the current question as input prompt. For HotpotQA, however, only the story and the current question is given as input prompt. We calculated the EM and F1 score for the two models on the TS-R and TS-R+Aug datasets for CoQA and the OS-R and OS-R+Aug datasets for HotpotQA. Table 6 shows these scores. For CoQA, we see that *text-davinci-002* achieves very high scores; for nearly $\sim$ 50% of cases the model continues to predict the ground truth answer for TS-R. This pathological behavior is similar to other models studied in this work. While *text-davinci-003* does achieve lower scores for TS-R, it sill predicts the ground truth answer for $\sim$ 30% of the cases. The scores for both models are lower for TS-R+Aug than TS-R. Unlike CoQA, for HotpotQA, *text-davinci-002* achieves lower scores than *text-davinci-003* on OS-R. We see that *text-davinci-003* predicts ground truth answer for 25.6% of the cases of OS-R. The scores for both the models are lower on OS-R+Aug than OS-R. Overall, these result show that

4 `https://platform.openai.com/docs/models/gpt-3-5`.

Table 5
EM, F1 score, and percentage of unknown predictions (i.e., *unk*%) of the models under the two training strategies for HotpotQA.

Model	Dataset	OT			IBT		
		F1	EM	unk%	F1	EM	unk%
BERT-base	OS	72.3	56.7	0.16	71.2	55.8	1.00
	OS-R	59.5	48.5	4.60	0.4	0.3	99.05
	OS-R+Aug	63.1	52.2	4.27	3.1	2.2	93.96
BERT-large	OS	74.8	59.6	0.21	73.8	58.5	1.24
	OS-R	63.7	53.8	5.63	0.5	0.4	99.17
	OS-R+Aug	64.7	54.2	5.36	2.63	2.05	95.46
RoBERTa-base	OS	72.0	56.7	0.16	72.7	57.4	0.73
	OS-R	66.2	59.1	0.86	0.6	0.5	98.86
	OS-R+Aug	36.9	15.7	0.94	0.9	0.6	97.93
RoBERTa-large	OS	80.0	64.5	0.18	79.7	64.4	0.70
	OS-R	75.2	70.0	2.84	0.6	0.5	99.06
	OS-R+Aug	40.3	18.4	3.90	0.9	0.6	97.86
XLNet-base	OS	74.2	60.1	0.07	73.5	59.4	1.05
	OS-R	63.0	53.0	0.73	0.6	0.4	98.85
	OS-R+Aug	63.5	53.9	1.16	3.1	2.4	94.30
XLNet-large	OS	80.0	66.1	0.23	77.4	63.5	1.03
	OS-R	68.5	59.1	0.60	0.4	0.3	99.21
	OS-R+Aug	62.7	53.7	9.12	1.6	1.1	96.81

Table 6
Deletion intervention: EM and F1 score for the two InstructGPT models.

Model	Dataset		EM	F1
text-davinci-002	CoQA	TS-R	46.4	58.5
		TS-R+Aug	40.4	53.3
	HotpotQA	OS-R	14.1	32.2
		OS-R+Aug	11.4	29.5
text-davinci-003	CoQA	TS-R	28.0	45.6
		TS-R+Aug	17.3	34.9
	HotpotQA	OS-R	25.6	41.7
		OS-R+Aug	18.0	34.1

InstructGPT models do not respond appropriately to deletion intervention. We provide several examples of deletion intervention in Section 10.

5.1 Intervention-based Training

To enhance the sensitivity of the language models to deletion intervention, we propose a simple intervention-based training (IBT). In this training strategy, we train the model on multiple datasets simultaneously. For CoQA, the model is trained to predict the ground truth answer for OS and TS, whereas for TS-R, the model is trained to predict *unknown*. Similarly, for HotpotQA, the model is trained to predict the ground truth answer for OS and to predict *unknown* for OS-R. Note that the models are trained for the same number of epochs under both the training strategies. The IBT in Tables 4 and 5 shows EM (exact match), F1, and the percentage of unknown predictions (unk%) of

the models for this training strategy on CoQA and HotpotQA, respectively. Firstly, we observe that for the datasets OS and TS of CoQA and OS of HotpotQA, the training strategy IBT is on par with the strategy OT. Also, we see that the performance of the models drops significantly and all the models also have very high unk% ($>$ 90%) post deletion intervention (TS-R for CoQA and OS-R for HotpotQA). Thus, IBT is able to make the models highly sensitive to deletion intervention. Furthermore, for OS-R+Aug of HotpotQA, the models' performance is similar to OS-R under IBT. For TS-R+Aug of CoQA, RoBERTa and XLNet have very high unk%. While the unk% for BERT drops significantly, IBT still fares better than OT on TS-R+Aug dataset. This shows that the models trained under IBT do not solely rely on superficial cues from TS-R (or OS-R) to predict *unknown*.

5.2 In-depth Analysis of Intervention-based Training

In this section, we study the inner-workings of these models in order to explain the effectiveness of intervention-based training against deletion intervention. As mentioned in Section 4.2, CLS embedding plays a crucial role in predicting an answer to a particular question. Hence, to begin with, we look at the cosine similarity (*cossim*) between CLS embeddings of OS and TS under the two training strategies (OT and IBT). Similarly, we also look at the cossim between CLS embeddings of OS and TS-R under the two training strategies. Figures 1 and 2 show the histogram of cossim on the development

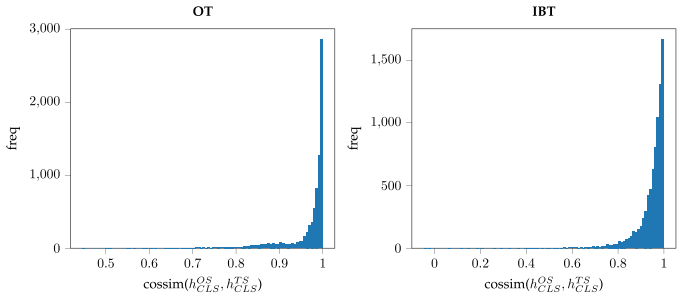

Figure 1
RoBERTa-large: Histogram plot of cosine similarity between CLS embedding for OS and TS under two training strategies (OT on left and IBT on right).

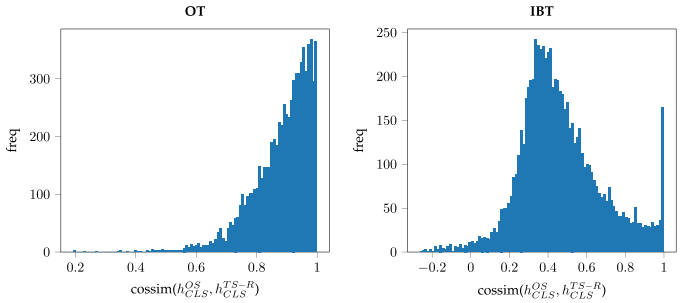

Figure 2
RoBERTa-large: Histogram plot of cosine similarity between CLS embedding for OS and TS-R for the two training strategies (OT on left and IBT on right).

set of CoQA for RoBERTa-large. In Figure 1, we see that the two histograms follow a similar pattern. The cossim is very high for almost all the cases. This is interesting since it shows that even if a significant chunk is removed from the story, it doesn't affect the CLS embedding in any meaningful way. However, in Figure 2, there is a drastic difference between the two histograms. Whereas the histogram for the OT strategy still follows a similar pattern as before, the histogram for IBT shows a significant drop in cossim. This shows that, for most of the cases under IBT, the CLS embedding is heavily affected once the rationale is removed from the story.

This effect is not only local to the CLS token but rather is observed for all the input tokens. To show this, we look at the cosine similarity of common tokens of OS and TS under the two training strategies, and similarly, the cosine similarity of common tokens of OS and TS-R under the two training strategies. Figures 3 and 4 show the corresponding histogram for RoBERTa-large. Here also, we can see that the cosine similarity of common tokens in OS and TS is very high for both training strategies. Once again, the model's representation of the common words doesn't seem to be affected by the removal of large parts of the textual context; this indicates either that the model finds the larger context irrelevant to the task or it might not be capable of encoding long-distance contextual information for this task.

For common tokens in OS and TS-R, however, there is a stark contrast between the two training strategies. For OT, the cosine similarity of common words still remain high but for IBT, the cosine similarity drops by a large margin. This shows that, under

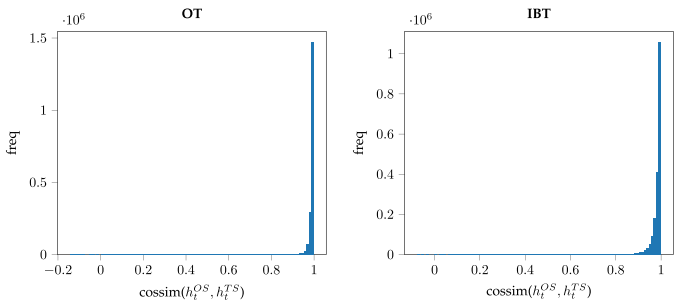

Figure 3
RoBERTa-large: Histogram plot of cosine similarity between common tokens of OS and TS for the two training strategies (OT on left and IBT on right).

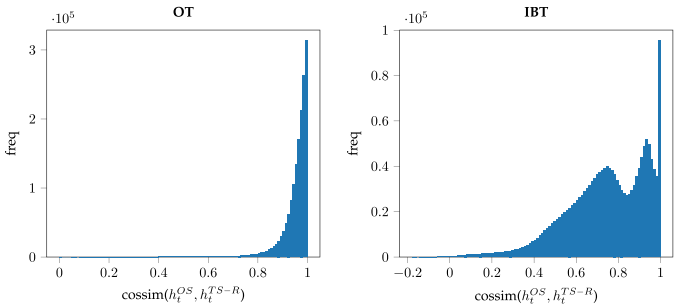

Figure 4
RoBERTa-large: Histogram plot of cosine similarity between common tokens of OS and TS-R under two training strategies (OT on left and IBT on right).

Table 7
Cosine similarity (mean ± std) of CLS embeddings for the two strategies.

Model	OT		IBT	
	cossim($h^{OS}_{CLS}, h^{TS}_{CLS}$)	cossim($h^{OS}_{CLS}, h^{TS-R}_{CLS}$)	cossim($h^{OS}_{CLS}, h^{TS}_{CLS}$)	cossim($h^{OS}_{CLS}, h^{TS-R}_{CLS}$)
BERT-base	0.99 ± 0.02	0.97 ± 0.03	0.96 ± 0.07	0.33 ± 0.34
BERT-large	0.99 ± 0.02	0.98 ± 0.04	0.95 ± 0.12	0.42 ± 0.31
RoBERTa-base	0.96 ± 0.04	0.92 ± 0.08	0.94 ± 0.06	0.55 ± 0.22
RoBERTa-large	0.97 ± 0.06	0.88 ± 0.10	0.95 ± 0.07	0.47 ± 0.21
XLNet-base	0.99 ± 0.03	0.95 ± 0.09	0.97 ± 0.06	0.27 ± 0.33
XLNet-large	0.98 ± 0.04	0.89 ± 0.15	0.98 ± 0.05	0.53 ± 0.21

Table 8
Cosine similarity (mean ± std) of common tokens for the two strategies.

Model	OT		IBT	
	cossim(h^{OS}_{t}, h^{TS}_{t})	cossim(h^{OS}_{t}, h^{TS-R}_{t})	cossim(h^{OS}_{t}, h^{TS}_{t})	cossim(h^{OS}_{t}, h^{TS-R}_{t})
BERT-base	0.99 ± 0.04	0.94 ± 0.07	0.96 ± 0.06	0.66 ± 0.22
BERT-large	0.99 ± 0.04	0.94 ± 0.06	0.98 ± 0.04	0.71 ± 0.21
RoBERTa-base	0.99 ± 0.04	0.95 ± 0.06	0.97 ± 0.05	0.75 ± 0.20
RoBERTa-large	0.99 ± 0.03	0.95 ± 0.06	0.98 ± 0.04	0.74 ± 0.19
XLNet-base	0.96 ± 0.08	0.90 ± 0.13	0.96 ± 0.08	0.57 ± 0.40
XLNet-large	0.94 ± 0.14	0.86 ± 0.24	0.94 ± 0.15	0.52 ± 0.44

Table 9
Cosine similarity (mean ± std) of CLS and common token embeddings for HotpotQA.

Model	OT		IBT	
	cossim($h^{OS}_{CLS}, h^{OS-R}_{CLS}$)	cossim(h^{OS}_{t}, h^{OS-R}_{t})	cossim($h^{OS}_{CLS}, h^{OS-R}_{CLS}$)	cossim(h^{OS}_{t}, h^{OS-R}_{t})
BERT-base	0.86 ± 0.12	0.85 ± 0.15	-0.03 ± 0.20	0.64 ± 0.20
BERT-large	0.53 ± 0.42	0.75 ± 0.21	0.17 ± 0.14	0.55 ± 0.18
RoBERTa-base	0.91 ± 0.07	0.84 ± 0.13	0.32 ± 0.22	0.59 ± 0.18
RoBERTa-large	0.92 ± 0.11	0.79 ± 0.20	0.40 ± 0.17	0.54 ± 0.16
XLNet-base	0.96 ± 0.05	0.92 ± 0.15	0.00 ± 0.23	0.87 ± 0.20
XLNet-large	0.81 ± 0.25	0.91 ± 0.16	-0.01 ± 0.18	0.77 ± 0.29

IBT, the embeddings of the input tokens are more contextualized with respect to the rationale. Due to this, under IBT, the word embeddings become significantly altered once the rationale is removed from the story. Similar to RoBERTa-large, other models also exhibit similar pattern of cosine similarity for CLS and common tokens, as shown in Tables 7 and 8 for CoQA, and Table 9 for HotpotQA. From Table 7, we can see that, for all the models, cossim($h^{OS}_{CLS}, h^{TS-R}_{CLS}$) for IBT is much lower than the corresponding cosine similarity for OT; whereas cossim($h^{OS}_{CLS}, h^{TS}_{CLS}$) is similar for both the strategies. Similarly, Table 8 shows that, for all the models, cossim(h^{OS}_{t}, h^{TS-R}_{t}) for IBT is much lower than the corresponding cosine similarity for OT; whereas cossim(h^{OS}_{t}, h^{TS}_{t}) is similar for both the strategies. For HotpotQA, Table 9 shows that IBT has lower cossim($h^{OS}_{CLS}, h^{OS-R}_{CLS}$) and cossim($h^{OS}_{t}, h^{OS-R}_{t}$) for all the models in comparison to OT.

From a more conceptual perspective, the sensitivity to the rationale in IBT suggests that IBT is providing the kind of instances needed to confirm the counterfactual, *were the rationale not present, the model would not answer as it does when the rationale is present*. Thus, at a macro level, attention-based models can locate spans of text crucial to determining semantic content through particular forms of training.

6. Negation Intervention

In this section, we detail our experiments on negation intervention. Negation intervention investigates possible causal dependencies of a model's inferences based on logical structure, in particular the scope of negation operators. As we stated in Section 2, the idea behind negation intervention is to alter a text with an intervention n such that $T, Q \models \psi$ iff $n(T), Q \models \neg\psi$.

For negation intervention, we randomly sampled 275 yes-no questions for CoQA. We appropriately modified the rationale in the truncated story (i.e., TS) for these samples in order to switch the answer from yes to no and vice versa. The development set of HotpotQA contained 458 yes-no questions. Out of these, we randomly sampled 159 questions. Almost all the questions followed the format *"Were both X and Y Z?"* (e.g., "Were Scott Derrickson and Ed Wood of the same nationality?"). For these samples, we modify the rationale in the original story (i.e., OS) in order to switch the answer. Table 10 shows the effect of negation intervention for the model trained under OT for the two datasets. In the table, Org-Acc refers to accuracy of the model on the original sample, Mod-Acc refers to accuracy of the model post negation intervention (i.e., with respect to the modified ground truth answer), and Comb-Acc refers to the percentage of cases where the model answered correctly for both original and modified sample. Table 10 shows a $\sim 20\%$ drop in accuracy for CoQA for all the models when we compare Org-Acc and Mod-Acc. For HotpotQA, the drop in accuracy is even higher for BERT and XLNet. This significant drop highlights the inability of the models to handle negation intervention. The low Comb-Acc scores of the models across the two datasets further highlight this fact. Switching to the IBT regime provided no significant difference. This indicates that another type of training will be needed for these models to take into systematic account the semantic contributions of negation.

A natural option is to train over negated examples and non-negated examples. Kassner and Schütze (2020) perform such an experiment and conclude that transformer-style models do not learn the meaning of negation. And Hosseini et al. (2021) provide a particular training regime that seems to improve language models' performance on the dataset of negated examples introduced by Kassner and Schütze (2020).

Table 10
Effect of negation intervention on different models.

Model	CoQA			HotpotQA		
	Org-Acc	Mod-Acc	Comb-Acc	Org-Acc	Mod-Acc	Comb-Acc
BERT-base	78.2	58.9	41.5	68.6	39.0	8.8
BERT-large	84.7	65.1	52.0	71.1	39.0	10.1
RoBERTa-base	81.8	61.8	47.6	61.6	45.9	9.4
RoBERTa-large	94.2	72.7	67.3	84.9	62.3	49.1
XLNet-base	85.1	64.7	52.0	69.8	37.1	6.9
XLNet-large	90.2	68.7	59.6	84.3	47.8	33.3

Nevertheless, while Kassner and Schütze's (2020) conclusion is compatible with our findings, we are not sanguine that Hosseini et al.'s (2021) training regime will improve model performance on the operation of negation intervention. The interventions we needed to make to induce the appropriate shifts in answers often depended on quite important shifts in material. Simple insertions of negation often seemed to disrupt the coherence and flow of the text; these disruptions could provide superficial clues for shifting the model's behavior in a task. To give an example, here is a rationale from one of the stories in the CoQA dataset:

(2) A law enforcement official told The Dallas Morning News that a door was apparently kicked in

Given the question, *Was the house broken into?* (the original answer was *yes*), we changed the rationale to:

(3) A law enforcement official told The Dallas Morning News that a door was open, leaving the possibility that the killers had been invited in

to get a negative answer to the question.

In this intervention, we didn't insert a negation but rather changed the wording to obtain a text inconsistent with the original answer. More generally, in only 72 out of 275 cases (26%), we added or removed "no/not". And within these 72 cases, only around 25 cases featured the simple addition/removal of "no/not" (e.g., the replacement of *six corn plants* to *but no corn plants*). In the rest of the cases, although we added/removed "no/not", we made more substantive changes to the story. Here is one example, where the question was, *did the wife console the boy?* The original rationale was as follows:

(4) "Robert Meyers said his wife tried to help Nowsch. "My wife spent countless hours at that park consoling this boy," he said.

We changed this to:

(5) Robert Meyers said his wife did not try to help Nowsch. "My wife spent countless hours at that park tormenting this boy," he said.

For the other 203 out of 275 cases (74%), there were lexical changes and substantial changes to the rationale to preserve stylistic consistency.

Thus the simple addition/removal of "no/not" cases numbered around 25 cases ($\sim$ 10%). In general, the models were able to switch their answers on such cases. Out of 73 cases, where RoBERTa-large answered the question for the original story correctly and didn't switch the answer post negation intervention, there are only 6 trivial cases. For HotpotQA, given the format of the question, simply adding/removing "no/not" is insufficient. For all the 159 cases, switching the answer required manipulating the entities present in the story (e.g., changing *American* to *Australian*).

Table 11
Effect of negation intervention on the two InstructGPT models.

Model	Dataset	Org-Acc	Mod-Acc	Comb-Acc
text-davinci-002	CoQA	88.0	61.1	53.5
	HotpotQA	79.2	49.7	31.4
text-davinci-003	CoQA	94.2	61.5	56.7
	HotpotQA	90.0	59.7	50.9

Similar to deletion intervention, we also test InstructGPT on our negation intervention dataset. Table 11 shows the results. From the tables, we can see that both *text-davinci-002* and *text-davinci-003* suffer a $\sim$ 20% drop in accuracy for CoQA and $\sim$ 30% drop in accuracy for HotpotQA. Thus, similar to other models, InstructGPT fails to respond well to negation intervention. The inferences involved in negation intervention are thus quite complex and go beyond the recognition of a simple negation. A mastery of such inferences would indicate a mastery not only of negation but of inconsistency, which would be a considerable achievement for a machine learned model. So simply alerting the model to the presence of negation will not suffice to guarantee reasoning ability with negation. The alternative is to create a corpus with many more negation intervention examples. However, it was difficult to construct the requisite data so as to meet our view of negation intervention since fine-tuning requires a lot more examples. We provide several examples of our negation intervention dataset for CoQA and HotpotQA in Section 10.

7. Predicate–Argument Structure

In this section, we study whether the models stay faithful to simple cases of predicate–argument structure. As we already mentioned in the Introduction, we propose two types of experiments. In the first simple experiment, we ask a question Q about the properties of objects in a text T. Given an answer ψ such that $T, Q \mid\models \psi$, we expect that for a semantically faithful model M_T that $M_T, Q \mid\models \psi$.

The second set of experiments is more involved. Formally, it involves the following set-up. Given:

- two questions, Q, Q',
- $T \models Q \leftrightarrow Q'$

we should have

$$T, Q \mid\models \psi \textit{ iff } T, Q' \mid\models \psi$$

To test for semantic faithfulness in these contexts, we devised synthetic, textual data for these experiments. We used five different schemas:

1. The *col1* car was standing in front of a *col2* house.

2. They played with a *col1* ball and *col2* bat.
3. The man was wearing a *col1* shirt and a *col2* jacket.
4. The house had a *col1* window and a *col2* door.
5. A *col1* glass was placed on a *col2* table.

where *col1* and *col2* denote two distinct colors. Using these 5 schemas and different color combinations, we constructed a dataset of 130 stories. For each story, we have 4 questions (two "yes" and two "no" questions). As an example, for the story, "The blue car was standing in front of a red house.", the two "yes" questions are "Was the car blue?" and "Was the house red?"; and two "no" questions are "Was the car red?" and "Was the house blue?". Thus, we have a total of 520 questions. The Org-Acc in Table 12 shows the accuracy of the models trained on CoQA on these questions and indicates a huge variance in accuracy across the models. We observed that BERT-base predicted *no* for all questions, and RoBERTa-base predicted *no* for most of the questions, while XLNet-base mostly predicted "yes". For the large models, RoBERTa-large and BERT-large achieved very high accuracies. We note, however, that this dataset is very simple.

These results indicate that the small models really didn't do much better than chance in answering our yes/no questions; hence either they didn't capture the predicate–argument structure of the sentences, or they could not use that information to reason to an answer to our questions. They failed on the most basic level. The large models fared much better, but this in itself didn't suffice to determine a causal link between the predicate–argument information and the inference to the answers.

Probing further, we then examined how the models fared under semantically equivalent questions. Q' is semantically the same as ($\equiv$) Q given context C iff they have the same answer sets in C (Bennett 1979; Karttunen 1977; Groenendijk 2003). In our situation, the context is given by the text T. Thus, we have $T \models Q \leftrightarrow Q'$ and $T, Q \mid\!\models \psi$ *iff* $T, Q' \mid\!\models \psi$. If M_T is semantically faithful and $T \models Q \leftrightarrow Q'$, then we should have $M_T, Q \mid\!\models \psi$ iff $M_T, Q' \mid\!\models \psi$. To construct semantically equivalent questions, we paraphrased the initial question in our dataset, namely, *"Was the car red?"* to *"Was there a red car?"*. This resulted in new set of 520 questions for the 130 stories. The Mod-Acc in Table 12 shows the accuracy of the models trained on CoQA on the modified questions. Apart from XLNet-base which predicted "yes" for most of the modified

Table 12
Effect of question paraphrasing on different models trained on CoQA. Org-Acc, and Mod-Acc denote accuracy on original and modified paraphrased question respectively. The number in bracket denote percentage of cases where the model predicted "no" as the answer. New-Acc refers to models' performance when "there" is replaced by "the" in the modified question.

Model	Org-Acc	Mod-Acc	New-Acc
BERT-base	50.0 (100.0)	69.4 (59.8)	50.0 (100.0)
BERT-large	95.2 (51.0)	77.3 (27.3)	98.1 (51.5)
RoBERTa-base	51.0 (99.0)	70.0 (78.5)	50.6 (99.4)
RoBERTa-large	99.4 (49.4)	95.0 (45.0)	99.8 (49.9)
XLNet-base	50.6 (6.0)	50.8 (0.7)	59.0 (13.3)
XLNet-large	75.2 (74.8)	79.8 (36.3)	78.7 (68.7)

questions and RoBERTa-large which retains it high accuracy, all the other models behave very differently from before. The accuracy of BERT-large drops drastically on these very simple questions, while BERT-base and RoBERTa-base perform significantly better on modified questions as they no longer mostly predict "no". For XLNet-large, while the two accuracies are similar, the model goes from predicting mostly "no" to mostly "yes". This contrast in behavior, as shown in Table 12, indicates that these models are unstable and lack semantic faithfulness on this task. There could be two possible reasons for this contrast, (i) different ordering of predicate and argument ("ball *col1*" vs. "*col1* ball"), (ii) different surface level words ("the" vs. "there"). We found the latter to be the case. To show this, we replace "there" with "the" in the modified question (i.e., "Was there a *col1* ball?" $\rightarrow$ "Was the a *col1* ball?"). The New-Acc in Table 12 shows the models' performance on this new question. We can see that the New-Acc is consistent with Org-Acc both in terms of accuracy and percentage of *no* predictions. This shows that the models are extremely sensitive to semantically unimportant words. For the models trained on HotpotQA, for both the question types (org and mod), we found that they either failed to recognize the question type as yes/no, or predicted "no". This is likely due to the fact that the yes/no questions in HotpotQA follow a fixed format, as discussed in Section 6, which is different from the one being used for this experiment.

Finally, we evaluate the performance of InstructGPT models. On the overall dataset of 1,040 questions (520 org questions and 520 mod questions), *text-davinci-002* achieved an accuracy of 96.7% (i.e., total of 34 failure cases), with Org-Acc of 99.6% and Mod-Acc of 93.8%. Interestingly, all the 34 failure cases in the original predicate–argument dataset were "yes" questions. For such cases, in many instances, we observed that adding an extra space to the prompt reverses the model's prediction. One such example is shown in Figure 5. As for *text-davinci-003*, the model achieved perfect accuracy. Unlike *text-davinci-002*, we found that *text-davinci-003* is stable in its prediction with regard to extra spaces in the prompt. However, there were 14 cases where *text-davinci-003* predicted "not necessarily" instead of "no". The question in all these cases was of the form "Was there a *col2* car?" Note that we had 26 cases with this question format and the model predicted "no" for the remaining 12 cases. This showcases instability in the model's prediction for two very similar input prompts.

We also tested the model's sensitivity to a contrastive set of examples in which we inserted negations in our predicate–argument dataset sentences (e.g., "The blue car was standing in front of a house that was not red." for the question "Was the house red?"). In contrast to its performance on negation intervention, the InstructGPT models achieved perfect accuracy on such negated examples. This further demonstrates that negation intervention is different from the tasks given by Naik et al. (2018), Kassner and Schütze (2020), Hossain et al. (2020), and Hosseini et al. (2021).

Figure 5
The *text-davinci-002* model predicts correctly when an extra space (shown in red) is added.

8. Discussion

In conclusion, concerning our findings about predicate–argument structure and logical structure more generally, we address three points.

1) Larger transformer-based models have shown to generally perform better than their smaller variants (Roberts, Raffel, and Shazeer 2020; Liu et al. 2019a). However, some exceptions to this trend have also been observed (Zhong et al. 2021). Our experiments show that with respect to the notion of semantic faithfulness, in general sensitivity to semantic structure and content, larger models fare better in predicate–argument experiments but not in our negation and deletion intervention experiments. For deletion intervention, they are mostly worse off than smaller models. Sections 5 and 6 show that InstructGPT also fails to tackle the two interventions in an efficient manner. For predicate–argument structure, the fact that the difference in models' behavior to two question types arises from irrelevant surface-level words is a big drawback since there are multiple ways to paraphrase a question. A possible way to tackle this issue is to translate the question into a logical form and train the model so that the contextualized embedding of the argument is more sensitive to its corresponding predicate. For example, the contextualized embedding of "car" is more sensitive to "blue" for the story "The blue car was standing in front of a red house." The sensitivity of a word w on another word w' can quantitatively be measured as the change in the contextualized embedding of the word w when the word w' is removed from its context. The higher the change, the higher the sensitivity. This sensitivity can then be used as a cue by the model to arrive at a prediction. We plan to explore this approach in the future.

2) Is prompting really superior to fine tuning? Our prompting experiment with InstructGPT allowed us to get results without fine-tuning. This is essentially zero-shot learning since no input–output pairs are provided in the prompt. However, for deletion and negation intervention, we observed that InstructGPT models do not present an advancement over other transformer-based models with respect to behavior post these interventions. Moreover, like Jiang et al. (2020), Liu et al. (2023), and Shin et al. (2021), we have found *text-davinci-002* to be extremely sensitive to what should be and intuitively is irrelevant information in the prompt. With regard to semantic faithfulness on predicate–argument structure, this shows an astonishing lack of robustness to totally irrelevant material, even if *text-davinci-002* scores very well on this dataset. This brittleness is telling: A semantically faithful model that exploits semantic structure to answer questions about which objects have which properties should not be sensitive to formatting changes in the prompt. This indicates to us that even if predicate–argument structure questions are answered correctly, *text-davinci-002* is not using that information as it should. *text-davinci-003* is stable to such insignificant changes in the prompt. However, the model still shows instability in its predictions for two very similar prompts, as highlighted earlier. Once again, we have our doubts that the right information (i.e., semantic structure) is being leveraged for the answer; if it were, *text-davinci-003* would answer in the same way for all the questions with "no" answers.

3) Extending semantic faithfulness beyond the question answering tasks in NLP. The definition of semantic faithfulness in Section 2 is geared to testing the semantic knowledge of LMs in question answering tasks. Question answering can take many forms and is a natural way to investigate many forms of inference or exploitations of semantic and logical structure. It also underlies many real-world NLP applications, like chatbots, virtual assistants, and Web searches (Liang et al. 2022). Semantic faithfulness can be extended to probe for a model's inferences concerning artificial languages like first order logic or any other formal language or programming language for which there

is a well defined notion of semantic consequence ($\models$). In such cases, the role of the "text" in semantic faithfulness would be played by a set of premises, a logical or mathematical theory, or code for an algorithm or procedure. Similar experiments of deletion or negation intervention could in principle be performed in these settings, which opens up a novel way of investigating LM models' performance on tasks like code generation (Chen et al. 2021; Sarsa et al. 2022). Alternatively, as suggested by Shin and Van Durme (2022), exploiting formal logical forms may help with semantic faithfulness.

9. Conclusion and Future Work

We have studied the semantic faithfulness of transformer-based language models for two intervention strategies—deletion intervention and negation intervention—and with respect to their responses to simple, semantically equivalent questions. Despite high performance on the original CoQA and HotpotQA, the models exhibited very low sensitivity to deletion intervention and suffered a significant drop in accuracy for negation intervention. They also exhibited unreliable and unstable behavior with respect to semantically equivalent questions (Q$\equiv$). Our simple intervention-based training strategy made the contextualized embeddings more sensitive to the rationale and corrected the models' erroneous reasoning in the case of deletion intervention.

Our research has exposed flaws in popular language models. In general, we have shown that even large models are not guaranteed to respect semantic faithfulness. This likely indicates that the models rely on superficial cues for answering questions about a given input text. While IBT is successful at remedying models' lack of attention to logical structure in cases of deletion intervention, it doesn't generalize well to the other experimental set-ups we have discussed. We do not have easy fixes for negation interventions or for the inferences involving predicate–argument structure. This is because it is difficult to generate enough data through negation intervention to retrain the model in the way we did for deletion intervention. Automating the process of negation intervention while preserving a text's discourse coherence and particular style remains a challenge. In addition, our investigations concerning predicate–argument structure and responses to semantically equivalent questions have pointed to a serious failing. We plan to explore the proposed approach in Section 8 to tackle this issue. Also, this work focused on a specific setting of semantic faithfulness where $\phi = \psi$ (refer to Equation 1). As part of future work, we plan to study the setting where multiple answers are possible.

Deletion and negation intervention modify the semantics of the input text. On the other hand, question equivalence experiments in this work preserve the semantics of the input. We plan to analyze the models' behavior to other possible semantic preserving interventions. A general solution to the problem of semantic unfaithfulness is something we have not provided in this article. However, we believe that key to solving this problem is a full scale integration of semantic structure without loss of inferential power in the transformer-based language models, something we plan to show in future work.

10. Appendix

Tables A.1, A.2, A.3, and A.4 show examples of deletion intervention for models trained under OT for CoQA and HotpotQA. Similarly, Tables A.5 and A.6 show examples of negation intervention for CoQA and HotpotQA, respectively. In all the examples presented here, the models' prediction remains unchanged post intervention, showing that the models rely on superficial cues for predicting an answer to a given question.

Table A.1
Deletion intervention examples from CoQA dataset. The rationale is marked in **bold** in TS. The models predict the same answer for TS and TS-R.

TS	My doorbell rings. On the step, I find the elderly Chinese lady, small and slight, holding the hand of a little boy. **In her other hand, she holds a paper carrier bag.**
TS-R	My doorbell rings. On the step, I find the elderly Chinese lady, small and slight, holding the hand of a little boy. paper carrier bag.
Conversation History	Who is at the door? An elderly Chinese lady and a little boy Is she carrying something? yes
Question	What?
Prediction	a paper carrier bag
	(a) BERT-base
TS	OCLC, currently incorporated as OCLC Online Computer Library Center, Incorporated, is an American nonprofit cooperative organization "dedicated to the public purposes of furthering access to the world's information and reducing information costs". **It was founded in 1967 as the Ohio College Library Center.**
TS-R	OCLC, currently incorporated as OCLC Online Computer Library Center, Incorporated, is an American nonprofit cooperative organization "dedicated to the public purposes of furthering access to the world's information and reducing information costs". 1967.
Conversation History	What is the main topic? OCLC What does it stand for? Online Computer Library Center
Question	When did it begin?
Prediction	1967
	(b) BERT-large
TS	Chapter XVIII "The Hound Restored" On the third day after his arrival at the camp Archie received orders to prepare to start with the hound, with the earl and a large party of men-at-arms, in search of Bruce. **A traitor had just come in and told them where Bruce had slept the night before.**
TS-R	Chapter XVIII "The Hound Restored" On the third day after his arrival at the camp Archie received orders to prepare to start with the hound, with the earl and a large party of men-at-arms, in search of Bruce. A traitor.
Conversation History	What was he told to start to do? Search for Bruce With what? with the hound, with the earl and a large party of men-at-arms
Question	Who gave them information about Bruce?
Prediction	A traitor
	(c) RoBERTa-base
TS	Can you imagine keeping an alien dog as a pet? This is what happens in CJ7 [...] When Ti falls off a building and dies, CJ7 saves his life. Because the dog loses all its power, it becomes a doll. **But Dicky still wears the dog around his neck.**
TS-R	Can you imagine keeping an alien dog as a pet? This is what happens in CJ7 [...] When Ti falls off a building and dies, CJ7 saves his life. Because the dog loses all its power, it becomes a doll. around his neck.
Conversation History	What did he become? a doll True or False: the boy loses the doll? False
Question	Where does he keep it, then?
Prediction	around his neck
	(d) RoBERTa-large

Table A.2
Deletion intervention examples from CoQA dataset. The rationale is marked in **bold** in TS. The models predict the same answer for TS and TS-R+Aug.

TS	Dhaka is the capital and largest city of Bangladesh. [...] At the height of its medieval glory, Dhaka was regarded as one of the wealthiest and most prosperous cities in the world. **It served as the capital of the Bengal province of the Mughal Empire twice (1608–39 and 1660–1704).**
TS-R+Aug	Dhaka is the capital and largest city of Bangladesh. [...] At the height of its medieval glory, Dhaka was regarded as one of the wealthiest and most prosperous cities in the world. I had to read the book twice to fully understand its theme.
Conversation History	Was it ever one of the wealthiest cities in the world? yes and when was that? At the height of its medieval glory
Question	How many times was it the capital of the Bengal province?
Prediction	twice
	(a) RoBERTa-base
TS	Andrew waited for his granddaddy to show up. They were going fishing. **His mom had packed them a lunch.**
TS-R+Aug	Andrew waited for his granddaddy to show up. They were going fishing. I packed them a lunch for their long road trip.
Conversation History	What was Andrew waiting for? His granddaddy Why? They were going fishing
Question	What did his mom do?
Prediction	packed them a lunch
	(b) RoBERTa-large
TS	ATLANTA, Georgia (CNN) – In 1989, the warnings were dire. **The Spike Lee film "Do the Right Thing" critics and columnists said, would provoke violence and disrupt race relations.**
TS-R+Aug	ATLANTA, Georgia (CNN) – In 1989, the warnings were dire. Spike Lee is a highly acclaimed filmmaker known for his innovative and thought-provoking films.
Conversation History	-
Question	Who created Do the Right Thing?
Prediction	Spike Lee
	(c) XLNet-base
TS	Once upon a time there was a cute brown puppy. He was a very happy puppy. His name was Rudy. Rudy had a best friend. His name was Thomas. Thomas had a nice dad named Rick. **Thomas and Rudy had been friends for almost a year.**
TS-R+Aug	Once upon a time there was a cute brown puppy. He was a very happy puppy. His name was Rudy. Rudy had a best friend. His name was Thomas. Thomas had a nice dad named Rick. I haven't seen my family in almost a year due to the pandemic.
Conversation History	Who was Rudy's best friend? Thomas
Question	How long have they been friends?
Prediction	almost a year
	(d) XLNet-large

Table A.3
Deletion intervention examples from HotpotQA dataset. The rationale is marked in **bold** in OS. The models predict the same answer for OS and OS-R.

OS	**Sergio Pérez Mendoza (born 26 January 1990) also known as "Checo" Pérez, is a Mexican racing driver, currently driving for Force India.** There have been six Formula One drivers from Mexico who have taken part in races since the championship began in 1950. **Pedro Rodríguez is the most successful Mexican driver being the only one to have won a grand prix. Sergio Pérez, the only other Mexican to finish on the podium, currently races with Sahara Force India F1 Team.**
OS-R	There have been six Formula One drivers from Mexico who have taken part in races since the championship began in 1950. Pedro Rodríguez
Question	Which other Mexican Formula One race car driver has held the podium besides the Force India driver born in 1990?
Prediction	Pedro Rodríguez

(a) BERT-base

OS	**The Manchester Terrier is a breed of dog of the smooth-haired terrier type. The Scotch Collie is a landrace breed of dog which originated from the highland regions of Scotland.** The breed consisted of both the long-haired (now known as Rough) Collie and the short-haired (now known as Smooth) Collie. It is generally believed to have descended from a variety of ancient herding dogs, some dating back to the Roman occupation, which may have included Roman Cattle Dogs, Native Celtic Dogs and Viking Herding Spitzes. **Other ancestors include the Gordon and Irish Setters.**
OS-R	The breed consisted of both the long-haired (now known as Rough) Collie and the short-haired (now known as Smooth) Collie. It is generally believed to have descended from a variety of ancient herding dogs, some dating back to the Roman occupation, which may have included Roman Cattle Dogs, Native Celtic Dogs and Viking Herding Spitzes. Scotch Collie
Question	Which dog's ancestors include Gordon and Irish Setters: the Manchester Terrier or the Scotch Collie?
Prediction	Scotch Collie

(b) BERT-large

OS	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).** Carrefour operates in more than 30 countries, in Europe,the Americas, Asia and Africa. Carrefour means "crossroads" and "public square" in French. The company is a component of the Euro Stoxx 50 stock market index. Euromarché ("Euromarket") was a French hypermarket chain. The first store opened in 1968 in Saint-Michel-sur-Orge. **In June 1991, the group was rebought by its rival, Carrefour, for 5,2 billion francs.**
OS-R	Carrefour S.A. is a French multinational retailer headquartered in Boulogne Billancourt, France, in the Hauts-de-Seine Department near Paris. Carrefour operates in more than 30 countries, in Europe,the Americas, Asia and Africa. Carrefour means "crossroads" and "public square" in French. The company is a component of the Euro Stoxx 50 stock market index. Euromarché ("Euromarket") was a French hypermarket chain. The first store opened in 1968 in Saint-Michel-sur-Orge. 1,462
Question	In 1991 Euromarché was bought by a chain that operated how any hypermarkets at the end of 2016?
Prediction	1,462

(c) RoBERTa-base

Table A.4
Deletion intervention examples from HotpotQA dataset. The rationale is marked in **bold** in OS. The models predict the same answer for OS and OS-R+Aug.

OS	**Marie Magdalene "Marlene" Dietrich (27 December 1901 – 6 May 1992) was a German actress and singer who held both German and American citizenship.** Throughout her unusually long career, which spanned from the 1910s to the 1980s, she maintained popularity by continually reinventing herself. **Marlene, also known in Germany as Marlene Dietrich-Porträt eines Mythos, is a 1984 documentary film made by Maximilian Schell about the legendary film star Marlene Dietrich.** It was made by Bayerischer Rundfunk (BR) and OKO-Film and released by Futura Film, Munich and Alive Films (USA).
OS-R+Aug	Throughout her unusually long career, which spanned from the 1910s to the 1980s, she maintained popularity by continually reinventing herself. It was made by Bayerischer Rundfunk (BR) and OKO-Film and released by Futura Film, Munich and Alive Films (USA). The year 1901 marked the beginning of a new century.
Question	The 1984 film "Marlene" is a documentary about an actress born in what year?
Prediction	1901
	(a) RoBERTa-large
OS	Current Mood is the third studio album by American country music singer Dustin Lynch. **It was released on September 8, 2017, via Broken Bow Records. The album includes the singles "Seein' Red" and "Small Town Boy", which have both reached number one on the Country Airplay chart.** "Small Town Boy" is a song recorded by American country music artist Dustin Lynch. **It was released to country radio on February 17, 2017 as the second single from his third studio album, "Current Mood".**
OS-R+Aug	Current Mood is the third studio album by American country music singer Dustin Lynch. "Small Town Boy" is a song recorded by American country music artist Dustin Lynch. September 8, 2017 was the day I graduated from college.
Question	When was the album that includes the song by Dustin Lynch released to country radio on February 17, 2017?
Prediction	September 8, 2017
	(b) XLNet-base
OS	**India Today is an Indian English-language fortnightly news magazine and news television channel. Aditya Puri is the Managing Director of HDFC Bank, India's largest private sector bank.** He assumed this position in September 1994, making him the longest-serving head of any private bank in the country. **India Today magazine ranked him 24th in India's 50 Most powerful people of 2017 list.**
OS-R+Aug	He assumed this position in September 1994, making him the longest-serving head of any private bank in the country. The employees are paid fortnightly instead of monthly.
Question	At what frequency the magazine publishes which ranked Aditya Puri 24th in India's 50 Most powerful people of 2017 list?
Prediction	fortnightly
	(c) XLNet-large

Table A.5
Negation intervention examples from CoQA dataset. The difference between the two stories is shown in **bold**. The models predict the same answer for original and modified stories.

Org. Story	Leeds is a city in West Yorkshire, England. [...] In the 17th and 18th centuries Leeds became a major centre for the production and trading of wool. During the Industrial Revolution, Leeds developed into a major mill town; **wool** was the dominant industry but flax, engineering, iron foundries, printing, and other industries were important.
Mod. Story	Leeds is a city in West Yorkshire, England. [...] In the 17th and 18th centuries Leeds became a major centre for the production and trading of wool. During the Industrial Revolution, Leeds developed into a major mill town; **timber** was the dominant industry but flax, engineering, iron foundries, printing, and other industries were important.
Conversation History	What part? West Yorkshire When did wool trade become popular? In the 17th and 18th centuries
Question	Was it the strongest industry?
Prediction	yes

(a) BERT-base

Org. Story	CHAPTER V–CLIPSTONE FRIENDS [...] Mr. Earl was **wifeless**, and the farm ladies heedless; but they were interrupted by Mysie running up to claim Miss Prescott for a game at croquet.
Mod. Story	CHAPTER V–CLIPSTONE FRIENDS [...] Mr. Earl was **married**, and the farm ladies heedless; but they were interrupted by Mysie running up to claim Miss Prescott for a game at croquet.
Conversation History	Who wants to take Miss Prescott from the conversation? Mysie To do what? game of croquet
Question	Is Mr. Earl married?
Prediction	no

(b) BERT-large

Org. Story	Once there was a beautiful fish named Asta. [...] Asta could not read the note. Sharkie could not read the note. They took the note to Asta's papa. "What does it say?" they asked. Asta's papa read the note. He told Asta and Sharkie, "This note is from a little girl. She wants to be your friend. If you want to be her friend, we can write a note to her. But you have to find another bottle so we can send it to her." **And that is what they did.**
Mod. Story	Once there was a beautiful fish named Asta. [...] Asta could not read the note. Sharkie could not read the note. They took the note to Asta's papa. "What does it say?" they asked. Asta's papa read the note. He told Asta and Sharkie, "This note is from a little girl. She wants to be your friend. If you want to be her friend, we can write a note to her. But you have to find another bottle so we can send it to her." **But they never found a suitable bottle.**
Conversation History	Who could read the note? Asta's papa What did they do with the note? unknown
Question	Did they write back?
Prediction	yes

(c) RoBERTa-base

Table A.6
Negation intervention examples from HotpotQA dataset. The difference between the two stories is shown in **bold**. The models predict the same answer for original and modified stories.

Org. Story	Hot Rod is a monthly American car magazine devoted to hot rodding, drag racing, and muscle cars modifying automobiles for performance and appearance. The Memory of Our People is a **magazine** published in the Argentine city of Rosario, a province of Santa Fe. The magazine was founded in 2004. Its original title in Spanish is "La Memoria de Nuestro Pueblo".
Mod. Story	Hot Rod is a monthly American car magazine devoted to hot rodding, drag racing, and muscle cars modifying automobiles for performance and appearance. The Memory of Our People is a **book** published in the Argentine city of Rosario, a province of Santa Fe. The book was founded in 2004. Its original title in Spanish is "La Memoria de Nuestro Pueblo".
Question	Are Hot Rod and The Memory of Our People both magazines?
Prediction	yes

(a) RoBERTa-large

Org. Story	Agee is a 1980 American documentary film directed by Ross Spears, about the writer James Agee. It was nominated for an Academy Award for Best Documentary Feature. To Shoot an Elephant is a 2009 documentary film about the 2008-2009 Gaza War directed by Alberto Arce and Mohammad Rujailahk.
Mod. Story	Agee is a 1980 American documentary film directed by Ross Spears, about the writer James Agee **and his war with the US**. It was nominated for an Academy Award for Best Documentary Feature. To Shoot an Elephant is a 2009 documentary film about the 2008-2009 Gaza War directed by Alberto Arce and Mohammad Rujailahk.
Question	Are Agee and To Shoot an Elephant both documentaries about war?
Prediction	yes

(b) XLNet-base

Org. Story	William Kronick is an American film and television writer, director and producer. He worked in the film industry from 1960 to 2000, when he segued into writing novels. Jonathan Charles Turteltaub (born August 8, 1963) is **an American film director and producer**.
Mod. Story	William Kronick is an American film and television writer, director and producer. He worked in the film industry from 1960 to 2000, when he segued into writing novels. Jonathan Charles Turteltaub (born August 8, 1963) is **a German television film director and writer**.
Question	Are William Kronick and Jon Turteltaub both television writers?
Prediction	no

(c) XLNet-large

Acknowledgments

For financial support, we thank the National Interdisciplinary Artificial Intelligence Institute ANITI (Artificial and Natural Intelligence Toulouse Institute), funded by the French 'Investing for the Future– PIA3' program under grant agreement ANR-19-PI3A-000. We also thank the projects COCOBOTS (ANR-21-FAI2-0005) and DISCUTER (ANR-21-ASIA-0005), and the COCOPIL "Graine" project of the Région Occitanie of France. This research is also supported by the Indo-French Centre for the Promotion of Advanced Research (IFCPAR/CEFIPRA) through project no. 6702-2 and Science and Engineering Research Board (SERB), Dept. of Science and Technology (DST), Govt. of India through grant file no. SPR/2020/000495.

References

Asher, Nicholas. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers. https://doi.org/10.1007/978-94-011-1715-9

Asher, Nicholas. 2011. *Lexical Meaning in Context: A Web of Words*. Cambridge University Press. https://doi.org/10.1017/CBO9780511793936

Asher, Nicholas, Soumya Paul, and Chris Russell. 2021. Fair and adequate explanations. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 79–97. https://doi.org/10.1007/978-3-030-84060-0_6

Asher, Nicholas, Soumya Paul, and Antoine Venant. 2017. Message exchange games in strategic contexts. *Journal of Philosophical Logic*, 46(4):355–404. https://doi.org/10.1007/s10992-016-9402-1

Asher, Nicholas and Sylvain Pogodalla. 2010. SDRT and continuation semantics. In *JSAI International Symposium on Artificial Intelligence*, pages 3–15. https://doi.org/10.1007/978-3-642-25655-4_2

Balasubramanian, Sriram, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi. 2020. What's in a name? Are BERT named entity representations just as good for any other name? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 205–214. https://doi.org/10.18653/v1/2020.repl4nlp-1.24

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org

Barwise, Jon and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219. https://doi.org/10.1007/BF00350139

Belinkov, Yonatan and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Bennett, M. 1979. *Questions in Montague Grammar*. Indiana University Linguistics Club.

Black, Sidney, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136. https://doi.org/10.18653/v1/2022.bigscience-1.9

Castelvecchi, D. 2022. Are chatGPT and alphacode going to replace programmers? *Nature News*. https://doi.org/10.1038/d41586-022-04383-z

Chang, Chen Chung and H. Jerome Keisler. 1990. *Model theory*. Elsevier.

Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chi, Ethan A., John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577. https://doi.org/10.18653/v1/2020.acl-main.493

Conia, Simone and Roberto Navigli. 2020. Bridging the gap in multilingual semantic role labeling: A language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410. https://doi.org/10.18653/v1/2020.coling-main.120

Conia, Simone and Roberto Navigli. 2022. Probing for predicate argument structures in pretrained language models. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4622–4632. https://doi.org/10.18653/v1/2022.acl-long.316

Davidson, Donald. 1967. Truth and meaning. *Synthese*, 17:304–323. https://doi.org/10.1007/BF00485035

De Groote, Philippe. 2006. Towards a Montagovian account of dynamics. In *Semantics and Linguistic Theory*, volume 16, pages 1–16. https://doi.org/10.3765/salt.v16i0.2952

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

Dowty, David R., Robert Wall, and Stanley Peters. 1981. *Introduction to Montague Semantics*. Synthese Library vol. 11. Springer. https://doi.org/10.1007/978-94-009-9065-4

Elazar, Yanai, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021a. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031. https://doi.org/10.1162/tacl_a_00410

Elazar, Yanai, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021b. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175. https://doi.org/10.1162/tacl_a_00359

Fernando, Tim. 2004. A finite-state approach to events in natural language semantics. *Journal of Logic and Computation*, 14(1):79–92. https://doi.org/10.1093/logcom/14.1.79

Fernando, Tim. 2022. Strings from neurons to language. In *Proceedings of the 3rd Natural Logic Meets Machine Learning Workshop (NALOMA III)*, pages 1–10.

Gardner, Matt, 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*, pages 1307–1323. https://doi.org/10.18653/v1/2020.findings-emnlp.117

Geva, Mor, Uri Katz, Aviv Ben-Arie, and Jonathan Berant. 2021. What's in your head? Emergent behaviour in multi-task transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8201–8215. https://doi.org/10.18653/v1/2021.emnlp-main.646

Graf, Thomas. 2019. A subregular bound on the complexity of lexical quantifiers. In *Proceedings of the 22nd Amsterdam Colloquium*, pages 455–464.

Groenendijk, Jeroen. 2003. Questions and answers: Semantics and logic. In the *2nd CologNET-ElSNET Symposium. Questions and Answers: Theoretical and Applied Perspectives*, pages 16–23.

Hewitt, John and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743. https://doi.org/10.18653/v1/D19-1275

Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138. https://doi.org/10.18653/v1/N19-1419

Hossain, Md Mosharaf, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118. https://doi.org/10.18653/v1/2020.emnlp-main.732

Hosseini, Arian, Siva Reddy, Dzmitry Bahdanau, R. Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 1301–1312. https://doi.org/10.18653/v1/2021.naacl-main.102

Hu, Minghao, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1596–1606. https://doi.org/10.18653/v1/D19-1170

Jia, Robin and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. https://doi.org/10.18653/v1/D17-1215

Jiang, Zhengbao, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438. https://doi.org/10.1162/tacl_a_00324

Ju, Ying, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on conversational question answering. *arXiv preprint 1909.10772*.

Kamp, H. and U. Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers.

Karttunen, L. 1977. Syntax and semantics of questions. *Linguistics and Philosophy*, 1(1):3–44. https://doi.org/10.1007/BF00351935

Kassner, Nora and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818. https://doi.org/10.18653/v1/2020.acl-main.698

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

Kusner, Matt J., Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094. https://doi.org/10.18653/v1/N19-1112

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):Article 5. https://doi.org/10.1145/3560815

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Naik, Aakanksha, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Pimentel, Tiago, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, pages 4609–4622. https://doi.org/10.18653/v1/2020.acl-main.420

Reddy, Siva, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266. https://doi.org/10.1162/tacl_a_00266

Reynolds, John C. 1974. On the relation between direct and continuation semantics. In *International Colloquium on Automata, Languages and Programming*, pages 141–156. https://doi.org/10.1007/978-3-662-21545-6_10

Roberts, Adam, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426. https://doi.org/10.18653/v1/2020.emnlp-main.437

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349

Sarsa, Sami, Paul Denny, Arto Hellas, and Juho Leinonen. 2022. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 27–43. https://doi.org/10.1145/3501385.3543957

Schuff, Hendrik, Heike Adel, and Ngoc Thang Vu. 2020. F1 is not enough! Models and evaluation towards user-centered explainable question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7076–7095. https://doi.org/10.18653/v1/2020.emnlp-main.575

Schölkopf, Bernhard. 2019. Causality for machine learning. *arXiv preprint arXiv:1911.10500*.

Shin, Richard, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715. https://doi.org/10.18653/v1/2021.emnlp-main.608

Shin, Richard and Benjamin Van Durme. 2022. Few-shot semantic parsing with language models trained on code. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5417–5425. https://doi.org/10.18653/v1/2022.naacl-main.396

Song, Liwei, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. 2021. Universal adversarial attacks with natural triggers for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3724–3733. https://doi.org/10.18653/v1/2021.naacl-main.291

Sun, Lichao, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip S. Yu, and Caiming Xiong. 2020. Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT. *arXiv preprint arXiv: 2003.04985*.

Talmor, Alon, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758. https://doi.org/10.1162/tacl_a_00342

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. https://doi.org/10.18653/v1/P19-1452

Tenney, Ian, 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 *International Conference on Learning Representations*.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing: System Demonstrations, pages 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. https://papers.nips.cc/paper_files/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

Yang, Zhilin, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. https://doi.org/10.18653/v1/D18-1259

Zhang, Wei Emma, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41. https://doi.org/10.1145/3374217

Zhong, Ruiqi, Dhruba Ghosh, Dan Klein, and Jacob Steinhardt. 2021. Are larger pretrained language models uniformly better? Comparing performance at the instance level. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3813–3827. https://doi.org/10.18653/v1/2021.findings-acl.334