Estimating the Causal Effects of Natural Logic Features in Transformer-Based NLI Models

Julia Rozanova[1], **Marco Valentino**[3], **André Freitas**[1,2,3]
[1]Department of Computer Science, University of Manchester, UK
[2] National Biomarker Centre, CRUK-MI, University of Manchester, UK
[3]Idiap Research Institute, Switzerland

Abstract

Rigorous evaluation of the causal effects of semantic features on language model predictions can be hard to achieve for natural language reasoning problems. However, this is such a desirable form of analysis from both an interpretability and model evaluation perspective, that it is valuable to investigate specific patterns of reasoning with enough structure and regularity to identify and quantify systematic reasoning failures in widely-used models. In this vein, we pick a portion of the NLI task for which an explicit causal diagram can be systematically constructed: the case where across two sentences (the premise and hypothesis), two related words/terms occur in a shared context. In this work, we apply causal effect estimation strategies to measure the effect of *context* interventions (whose effect on the entailment label is mediated by the semantic *monotonicity* characteristic) and interventions on the inserted word-pair (whose effect on the entailment label is mediated by the *relation* between these words). Extending related work on causal analysis of NLP models in different settings, we perform an extensive interventional study on the NLI task to investigate *robustness to irrelevant changes* and *sensitivity to impactful changes* of Transformers. The results strongly bolster the fact that similar benchmark accuracy scores may be observed for models that exhibit very different behaviour. Moreover, our methodology reinforces previously suspected biases from a causal perspective, including biases in favour of upward-monotone contexts and ignoring the effects of negation markers.

1. Introduction

There is an abundance of reported cases where high accuracies in NLP tasks can be attributed to simple heuristics and dataset artifacts (McCoy et al., 2019a). As such, when we expect a language model to capture a specific reasoning strategy or correctly use certain semantic features, it has become good practice to perform evaluations that provide a more granular and qualitative view into model behaviour and efficacy. In particular, there is a trend in recent work to incorporate causal measures and *interventional* experimental setups in order to better understand the captured features and reasoning mechanisms of NLP models (Vig et al., 2020; Finlayson et al., 2021; Stolfo et al., 2023; Geiger et al., 2021; Rozanova et al., 2023; Arakelyan et al., 2024).

In general, it can be hard to pinpoint all the intermediate features and critical representation elements which are guiding the inference behind an NLP task. However, in many cases there are subtasks which have enough semantic/logical regularity to perform stronger analyses and diagnose clear points of failure within larger tasks such as NLI and QA (Question Answering). As soon as we are able to draw a causal diagram which captures a portion of the model's expected reasoning capabillities, we may be guided in the design of interventional experiments which allow us to estimate causal quantities of interest, giving insight into how different aspects of the inputs are used by models.

In this work, we investigate a structured subset of the NLI task (Rozanova et al., 2022) to better understand the use of two semantic inference features by NLI models: concept relations and logical monotonicity. We use these intermediate abstracted semantic feature labels to construct *intervention sets* out of NLI examples which allow us to measure certain causal effects. Building upon recent work on causal analysis of NLP models Stolfo et al. (2023), we use the intervention sets to systematically and quantitatively characterise models' *sensitivity* to relevant changes in these semantic features and *robustness* to irrelevant changes.

Our contributions may be summarised as follows:

- Extending previous work on causal analysis of NLP models, we investigate a structured subproblem in NLI (in our case, a subtask based on natural logic (MacCartney and Manning, 2007)) and present a causal diagram which captures both desired and undesired potential reasoning routes which may describe model behaviour.
- We adapt the NLI-XY dataset of Rozanova et al. (2022) to a meaningful collection of *intervention sets* which enable the computation of certain causal effects.
- We calculate estimates for undesired direct

LREC-COLING 2024, pages 6319–6329
20-25 May, 2024.

causal effects and desired total causal effects, which also serve as a quantification of model robustness and sensitivity to our intermediate semantic features of interest.

- We compare a suite of BERT-like NLI models, identifying behavioural weaknesses in high-performing models and behavioural advantages in some worse-performing ones.

To the best of our knowledge, we are the first to complement previous observations of models' brittleness with respect to context monotonicity with the evidence of causal effect measures[1], as well as presenting new insights that over-reliance on lexical relations is consequently also tempered by the same improvement strategies.

2. Problem Formulation

2.1. A Structured NLI Subtask

As soon as we have a concrete description of how a reasoning problem *should* be treated, we can begin to evaluate how well a model emulates the expected behaviour and whether it is capturing the semantic abstractions at play.

In this work, we consider an NLI subtask which comes from the broader setting of *Natural Logic* (MacCartney and Manning, 2007; Hu and Moss, 2018; Sánchez, 1991). As it has a rigid and well-understood structure, it is often used in interpretability and explainability studies for NLI models (Geiger et al., 2021; Richardson et al., 2019a; Geiger et al., 2022; Rozanova et al., 2022, 2021). We begin with the format described in (Rozanova et al., 2022) (we refer to this work for more detailed description and full definitions).

Consider two terms/concepts with a known *relation label*, such as one of the pairs:

Word/Term x	Word/Term y	Relation
brown sugar	sugar	$x \sqsubseteq y$
mammal	lion	$x \sqsupseteq y$
computer	pomegranate	$x \# y$

Suppose the two terms occur in an identical context (comprising of a natural language sentence, like a template), for example:

Premise	I do not have any **sugar**.
Hypothesis	I do not have any **brown sugar**.

A semantic property of the natural language context called *monotonicity* determines whether there is an *entailment relation* between the sentences generated upon substitution/insertion of given related terms (formally, this is monotonicity in the sense of preserving the "order" between the inserted terms to an equally-directed entailment relation between the sentences.) The context monotonicity may either be *upward* (↑) or *downward* (↓, as in the example above) or *neither*.

M \ R	$\sqsubseteq$	$\sqsupseteq$	$\#$
↑	Entailment	Non-Entailment	Non-Entailment
↓	Non-Entailment	Entailment	Non-Entailment

Table 1: The entailment gold labels as a function of two semantic features: the context montonicity (M) and the relation (R) of the inserted word pair.

Variable	Description
G	Gold Label
C	Context
M	Context Monotonicity
W	Inserted Word Pair
R	Word-Pair Relation

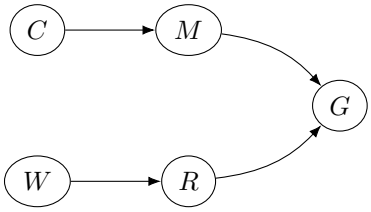


Figure 1: Causal Diagram for the Natural Logic Subtask

The effect of the context's monotonicity in conjunction with the relation between the inserted words on the gold entailment label is summarised in table 1. The authors of Rozanova et al. (2022) provide a thus-formatted dataset called *NLI-XY*, which we use as the basis for our causal effect estimation experiments.

Throughout the remainder of this paper, we will represent an NLI-XY example n as a tuple $n = (c, m, w, r, g)$ in which c is the shared natural language context, m is its monotonicity label, w is a pair (w_1, w_2) of nouns/noun phrases which will be inserted into the context (we refer to these as the *inserted word pair* for brevity), r is the concept inclusion relation label for w and g is the entailment gold label arising from m and r as per table 1. We denote by $P(Y \mid C = c, W = w)$ the probabilistic output of a trained NLI model with the example n as the NLI input (in particular, the input is the premise–hypothesis pair $(c(w_1), c(w_2))$).

As we have chosen a coarse segmentation of the monotonicity reasoning problem, we can present a simple causal diagram which illustrates our expectations for the correct reasoning scheme for a

[1] Our code is available at https://github.com/juliarozanova/counterfact_nli.

fixed class of NLI problems. The diagram in figure 1 shows the features on which the gold label is dependent on in the NLI-XY dataset: only the context monotonicity M and the concept pair relation R, which are respectively dependent on the content of the natural language context C and the concept pair / word pair W which is substituted into it. The exact values of the gold label with respect to these features may be referenced in table 1.

Naturally, it is always likely that models may fail to follow the described reasoning scheme for these NLI problems. In the next section (2.2), we propose a causal diagram which also captures the reasoning possibilities an NLI model may follow, accounting for possible confounding heuristics via unwanted direct effects.

2.2. The Causal Structure of Model Decision-Making

In an ideal situation, a strong NLI model would identify the word-pair relation and the context monotonicity as the abstract variables relevant to the final entailment label. In this case, these features would causally affect the model prediction in the same way they affect the gold label. Realistically, as shown in illuminating studies such as McCoy et al. (2019b), models identify unexpected biases in the dataset and may end up using accidental correlations output labels, such as the frequency of certain words in a corpus. For example, Richard T. McCoy (2019) demonstrate how models can successfully exploit the presence of negation markers to anticipate non-entailment, even when it is not semantically relevant to the output label.

To ensure that the semantic features themselves are taken account into the model's output and not other surface-level confounding variables, one would like to perform interventional studies which alter the value of the target feature but not other confounding variables. This is, in many cases, not feasible (although attempts are sometimes made to at least perform interventions that only make minimal changes to the textual surface form, as in Kaushik et al. (2020).)

Stolfo et al. (2023) argue that it is useful to quantify instead the direct impact of irrelevant surface changes (controlling for values of semantic variables of interest) and compare them to *total causal effects* of input-level changes: doing so, we may posit deductions about the flow of information via the semantic variables (or lack thereof). For analyses where there is an attempt to align intermediate variables with explicit internals, see Vig et al. (2020) and Finlayson et al. (2021) for a mediation analysis approach, or Geiger et al. (2021) for an alignment strategy based on causal abstraction theory.

Variable	Description
Y	Model Prediction
G	Gold Label
C	Context
M	Context Monotonicity
S	Context Textual Surface Form
W	Inserted Word Pair
R	Word-Pair Relation
T	Word-Pair Textual Surface Form

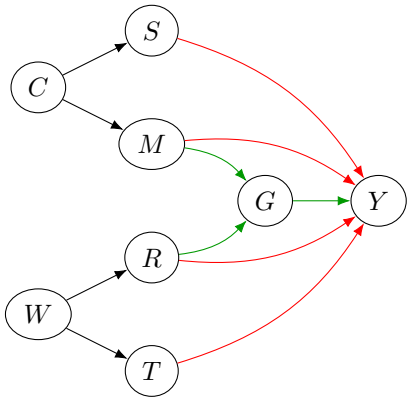


Figure 2: Specification of the causal diagram for possible routes of model reasoning for NLI-XY problems. Green edges indicate *desired* causal influence, while red edges indicate *undesired* paths of causal influence via surface-level heuristics.

Diagram Specification We follow Stolfo et al. (2023) in the strategy of explicitly modeling the "irrelevant surface form" of the input text portions as variables in the causal diagram. Their setting of *math word problems* is decomposed into two compositional inputs: a question template and two integer arguments. Our setting follows much the same structure: our natural language "context" plays the same role as their "template", but our arguments (an inserted word pair) have an additional layer of complexity as we also model the *relation* between the arguments as an intermediate reasoning variable rather than the values themselves (as such, the structure of their template modeling in their causal diagram is more applicable than the direct way they treat their numerical arguments.)

We present our own causal diagram in figure 2. We introduce the textual context C as an input variable, which is further decomposed into more abstract variables: its *monotonicity* M (which directly affects the gold truth G) and the textual surface form S of the context . The other input variable is the word-pair insertion which we will summarise as a single variable W. Once again, W has a potential effect on the model decision through its textual surface form T and via the relation R between the words. The gold truth G is dependent on M and R only. Finally, the outcome variable is the model

prediction Y. The paths for which we would like to observe the highest causal effect are the paths to Y from the inputs via M, R and through the gold truth variable G. However, each of S, T, M and R have direct links to the model output Y as well (indicated in red): these are potential direct effects which are *unwanted*. For example, we would not want a model to learn a prediction heuristic based directly on the variable M, such as consistently predicting non-entailment any time a downward monotone context is recognised. Similarly, a direct effect of S or T would look like a heuristic which predicts the entailment label purely based on the presence of words which happened to co-occur with that label in the training data. The key goal of this study is to compare the extent to which models exhibit the high causal effects for the *desired* diagram routes and lower causal effects for the *undesired* routes.

3. Estimating the Causal Effects

Given a fixed set N of NLI-XY examples, we define an *intervention* $\mathcal{I}$ on N as a set of pairs (n, n') of NLI-XY examples for (one for each $n \in N$), where $n' = (c', m', w', r', g')$ is a second NLI-XY example which represents a modified version of n (in practice, a modification of either c or w). We denote by N' the set of modified NLI-XY examples, so that $\mathcal{I} \subseteq N \times N'$.

For any pair $(n, n') \in \mathcal{I}$, we define the change-of-prediction indicator

$$CP(n, n') = \begin{cases} 1 & \text{if } y \neq y' \\ 0 & \text{if } y = y' \end{cases},$$

where

$$y = \arg\max_{i \in \{0,1\}} P(Y = i \mid C = c, W = w)$$

(namely, the model prediction which assigns the entailment label with the highest predicted probability) and

$$y' = \arg\max_{i \in \{0,1\}} P(Y = i \mid C = c', W = w').$$

Stolfo et al. (2023) refer to the average change-of-prediction quantity for a given intervention $\mathcal{I}$ as a *causal effect*. This causal effect quantity is named and interpreted differently depending on the conditions of the intervention: in particular, which variables are changed and which are kept constant throughout the intervention set over which we will take the average.

3.1. Interventions for Calculating TCE and DCE

The quantities of interest in Stolfo et al. (2023) are the *total causal effect* (TCE) of interventions on the variables which we would like to see having an effect on the prediction (in our case, C and W) and the *direct causal effect* (DCE) of interventions on the variables which we do *not* wish to unnecessarily impact the model prediction (in our case, T and S).

For a given *source* variable and *target* variable, whether we are measuring a DCE or TCE differs only in the design of the intervention set, which in turn depends on the structure of the causal diagram. For the design of the relevant intervention sets, we follow the strategy in Stolfo et al. (2023), as the upper portion of their causal diagram (concerning the natural language question template, its textual surface form and the implicit math operation) is equivalent to both the upper and lower half of our diagram in figure 2.

In this work, we provide four intervention sets: $\mathcal{I}_0, \mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$, each corresponding to the quantities TCE $(C$ on $Y)$, TCE $(W$ on $Y)$, DCE $(T \to Y)$ and DCE $(S \to Y)$ respectively. [2] We stick to their nomenclature of total causal effect (TCE) and direct causal effect (DCE), but define the quantities in the way that they are concretely calculated (in both our experiments and in Stolfo et al. (2023)): as an *estimate* of the causal effect quantity, which they present as an expected value of the change-of-prediction indicator.

(Desired) Total Causal Effects We estimate the total causal effect of the context C on the model prediction Y by constructing an intervention set $\mathcal{I}_0$ as follows: starting with a randomly sampled set N of NLI-XY examples, we intervene on each $n \in N$ by sampling a different context c' from the NLI-XY dataset which should result in a changed prediction, while keeping the inserted word pair w constant. In summary, every $(n, n') \in \mathcal{I}_0$ satisfies

$$(c \neq c', m \neq m', w = w', r = r', g \neq g').$$

We then calculate:

$$\text{TCE}(C \text{ on } Y) = \frac{1}{|\mathcal{I}_0|} \sum_{(n,n') \in I_0} CP(n, n')$$

Secondly, we estimate the total causal effect of the inserted word pair W on the model prediction Y by constructing an intervention set $\mathcal{I}_1$ as follows: starting with a randomly sampled set N of NLI-XY examples, we intervene on each $n \in N$ by sampling a different inserted word pair w' from the NLI-XY dataset which should result in a changed

[2] To be consistent with the notation in Stolfo et al. (2023), we will stylize these quantities as (for example) TCE(C on Y) and DCE($S \to Y$), where the arrow emphasizes that the quantity is specific to a direct path in the causal diagram (passing through no intermediate variables).

prediction, while keeping the shared context c constant. In summary, every $(n, n') \in \mathcal{I}_1$ satisfies

$$(c = c', m = m', w \neq w', r \neq r', g \neq g').$$

We then calculate:

$$\text{TCE}(W \text{ on } Y) = \frac{1}{|\mathcal{I}_1|} \sum_{(n,n') \in I_1} CP(n, n')$$

Following Stolfo et al. (2023), we interpret this quantity as a measure of model *sensitivity* to relevant context (respectively, inserted word pair) changes. As it quantifies how often the prediction changes when it should, we would like to see this value being as close to 1 as possible.

(Undesired) Direct Causal Effects The total causal effect does not distinguish whether this effect is mediated through the preferred causal route (for example, via context's monotonicity) or through a model heuristic based on the textual surface form: it is taking into account all possible routes of influence. The key suggestion in Stolfo et al. (2023) is that even though we have no feasible intervention strategies which allow us to calculate the causal effect of the intermediate variables M and R on Y as mediated through the gold label G (the effect of greatest interest to us), we may yield some insight into their causal influence by comparing the relevant TCE to the unwanted *direct causal effect* DCE ($S \rightarrow Y$) (respectively, DCE ($T \rightarrow Y$)).

To estimate the direct causal effect of the textual surface form S of the context C which is irrelevant to the context monotonicity M, we construct an intervention set $\mathcal{I}_2$ as follows: starting with a randomly sampled set N of NLI-XY examples, we intervene on each $n \in N$ by sampling a different context c' from the NLI-XY dataset while conditioning on the monotonicity (specifically, c' is chosen so that its monotonicity attribute m' is the same as that of c). The word pair w' (and therefore its relation r') are kept the same as in n, so the prediction is expected *not* to change. In summary, every $(n, n') \in \mathcal{I}_2$ satisfies

$$(c \neq c', m = m', w = w', r = r', g = g').$$

We then calculate:

$$\text{DCE}(S \rightarrow Y) = \frac{1}{|\mathcal{I}_2|} \sum_{(n,n') \in \mathcal{I}_2} CP(n, n')$$

To estimate the direct causal effect of the textual surface form T of the inserted word pair W which is irrelevant to the word pair relation R, we construct an intervention set $\mathcal{I}_3$ as follows: starting with a randomly sampled set N of NLI-XY examples, we intervene on each $n \in N$ by sampling a different inserted word pair w' from the NLI-XY dataset while conditioning on the word pair relation (specifically, w' is chosen so that its relation attribute r' is the same as that of w). The context c' (and therefore its monotonicity m') are kept the same as in n, so the prediction is expected *not* to change. In summary, every $(n, n') \in \mathcal{I}_3$ satisfies

$$(c = c', m = m', w \neq w', r = r', g = g').$$

We then calculate:

$$\text{DCE}(T \rightarrow Y) = \frac{1}{|\mathcal{I}_3|} \sum_{(n,n') \in \mathcal{I}_3} CP(n, n')$$

Once again following Stolfo et al. (2023), we interpret this quantity as a measure of model *robustness* to irrelevant context (respectively, inserted word pair) changes. As it quantifies how often the prediction changes in cases when it *shouldn't*, we would like to see this value being as close to 0 as possible. We present examples and dataset statistics for the intervention sets in the next section, along with the summary of the intervention schema in table 4.

4. Experimental Setup

4.1. Data and Interventions

We use the NLI-XY evaluation dataset to construct intervention pairs (n, n') by using a sampling/filtering strategy as in (Stolfo et al., 2023) according to the intervention schema in table 4. In particular, for constructing *context* interventions, we sample a seed set of 400 NLI-XY premise/hypothesis pairs. This is the *pre-intervention* NLI example. For each, we fix the insertion pair and filter through the NLI-XY dataset for all examples with the shared insertion pair but different context, conditioned as necessary on the properties of the other variables as in the intervention schema. For insertion pairs, we do the opposite. The number of interventions we produce in this way for our experiments are reflected in the last column of table 4. In summary, the changes are context replacements and related word-pair replacements; we provide text-level examples in tables 2 and 3 .

4.2. Model Choice and Benchmark Comparison

We include the following models [3] in our study: firstly, the models evaluated in NLI-XY pa-

[3]All pretrained models are from the Huggingface *transformers* library (Wolf et al., 2020), except for infobert and the pretrained model counterparts fine-tuned on HELP: their sources are linked in the README of the accompanying code.

Intervention Set	Target Quantity	Intervention Step	Premise	Hypothesis	M	R	G
$\mathcal{I}_1$	TCE(W on Y)	Before	There's a cat on the pc.	There's a cat on the machine.	$\uparrow$	$\sqsubset$	Entailment
		After	There's a cat on the tree.	There's a cat on the fruit tree.	$\uparrow$	$\sqsupset$	Non-Entailment
$\mathcal{I}_3$	DCE($T \rightarrow Y$)	Before	There are no students yet.	There are no first-year students yet.	$\downarrow$	$\sqsupset$	Entailment
		After	There are no people yet.	There are no women yet.	$\downarrow$	$\sqsupset$	Entailment

Table 2: Example word-pair insertion interventions for determining the total causal effect of label-relevant word-pair changes and the direct causal effect of label-irrelevant word-pair changes.

Intervention Set	Target Quantity	Intervention Step	Premise	Hypothesis	M	R	G
$\mathcal{I}_0$	TCE(C on Y)	Before	You can't live without fruit .	You can't live without strawberries .	$\uparrow$	$\sqsupset$	Non-Entailment
		After	All fruit study English.	All strawberries study English.	$\downarrow$	$\sqsupset$	Entailment
$\mathcal{I}_2$	DCE($S \rightarrow Y$)	Before	He has no interest in seafood .	He has no interest in oysters .	$\downarrow$	$\sqsupset$	Entailment
		After	I don't want to argue about this in front of seafood .	I don't want to argue about this in front of oysters .	$\downarrow$	$\sqsupset$	Entailment

Table 3: Example context interventions for determining the total causal effect of label-relevant context changes and the direct causal effect of label-irrelevant context changes.

Intervention Set	Target Measure	C	W	M	R	G	Interventions in Dataset
$\mathcal{I}_0$	TCE ($C \rightarrow Y$)	$\neq$	$=$	$\neq$	$=$	$\neq$	14270
$\mathcal{I}_1$	TCE ($W \rightarrow Y$)	$=$	$\neq$	$=$	$\neq$	$\neq$	22640
$\mathcal{I}_2$	DCE ($S \rightarrow Y$)	$\neq$	$=$	$=$	$=$	$=$	20910
$\mathcal{I}_3$	DCE ($T \rightarrow Y$)	$=$	$\neq$	$=$	$=$	$=$	25960

Table 4: Intervention schema and dataset statistics: which variables are held constant and which are changed in the construction of intervention sets for the calculation of the indicated effects.

per (Rozanova et al., 2022), namely roberta-large-mnli, facebook/bart-large-mnli, bert-base-uncased-snli and their counterparts fine-tuned on the HELP dataset (Yanaka et al., 2019b) Next, the infobert model, which is trained on three benchmark training sets of interest: MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015) and ANLI (Nie et al., 2020) (currently at the top of the leaderboard for the adversarial ANLI test set, as of January 2023) Lastly, another roberta-large checkpoint, also trained on all three benchmark NLI training training sets (as well as FEVER-NLI (Nie et al., 2019)). We report their scores on the mentioned benchmark datasets alongside the relevant total and direct causal effects we are interested in.

Note that as the HELP dataset is a two-class entailment dataset (as opposed to datasets like MNLI, which are three-class), we cannot directly compare existing reported scores. As such, we adapt the three-class scores to a two-class score by grouping two of the three-class labels ("contradiction" and "neutral") into the two-class umbrella label "non-entailment".

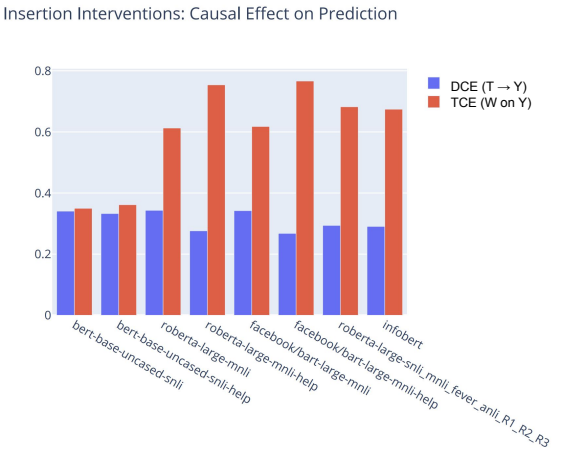
5. Results and Discussion

We examine and compare the results for the models listed in 4.2. We first observe the word-pair insertion intervention experiments in 5.1, then the context intervention experiments in 5.2 and finally present a categorical overview of these results in section 5.3, contextualised by benchmark scores.

5.1. Causal Effect of Inserted Word Pairs

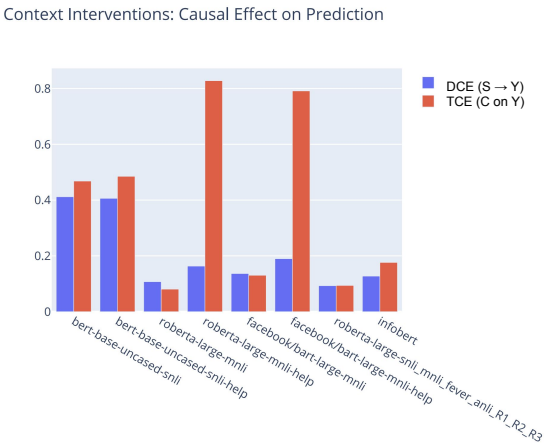
The results for the substituted word-pair intervention experiment are reported in figure 3. The most desireable outcome is a DCE($T \rightarrow Y$) which is *as low as possible* in combination with a TCE(W on Y) which is *as high as possible*. The lower this DCE, the higher the model robustness to *irrelevant word pair surface form* changes. On the other hand, the higher the specified TCE, the greater the model's sensitivity to *word pair insertion changes affecting the gold label*.

The largest delta between these two quantities can be seen in the roberta-large-mnli-help and facebook-bart-large-mnli-help models. This is important to note: the HELP dataset (Yanaka et al., 2019b) is explicitly designed to bolster model success on natural logic problems, but until now there has been little to no evidence that it improves the treatment of word-pair relations. In particular, the internal probing results in Rozanova et al. (2022) show that probing performance for the intermediate word-pair relation label decreases slightly for roberta-large-mnli after fine-tuning on HELP; as such, it was thought that the HELP improvements on natural logic could solely be attributed to improved context monotonicity treatment. Now, however, we observe distinct improvements in ro-



Model	DCE($T \rightarrow Y$)	TCE(W on Y)	TCE/DCE Ratio	Delta
bert-base-uncased-snli	0.341	0.350	1.027	0.009
bert-base-uncased-snli-help	0.332	0.361	1.087	0.029
roberta-large-mnli	0.343	0.613	1.785	0.269
roberta-large-mnli-help	0.276	0.754	2.730	0.478
facebook/bart-large-mnli	0.342	0.618	1.805	0.275
facebook/bart-large-mnli-help	0.268	0.766	2.863	0.499
roberta-large-snli_mnli_fever_anli_R1_R2_R3	0.294	0.682	2.321	0.388
infobert	0.291	0.674	2.320	0.384

Figure 3: Results for Insertion Interventions



Model	DCE($S \rightarrow Y$)	TCE(C on Y)	TCE/DCE Ratio	Delta
bert-base-uncased-snli	0.412	0.468	1.136	0.0563
bert-base-uncased-snli-help	0.406	0.485	1.194	0.079
roberta-large-mnli	0.107	0.081	0.751	-0.027
roberta-large-mnli-help	0.163	0.828	5.070	0.665
facebook/bart-large-mnli	0.136	0.130	0.954	-0.006
facebook/bart-large-mnli-help	0.189	0.791	4.167	0.601
roberta-large-snli_mnli_fever_anli_R1_R2_R3	0.093	0.093	1.008	0.001
infobert	0.127	0.176	1.385	0.049

Figure 4: Results for Context Interventions

bustness to irrelevant word-pair insertion changes and sensitivity to relevant ones.

More generally, the work in Rozanova et al. (2022) does indicate that the large MNLI-based models are already very successful in distinguishing the relation between substituted words. The word-pair relation label has a high *probing* result for all of these models, as well as strong signs of systematicity in their error analysis. This is in line with our observations of relatively large deltas between the DCE and TCE here, compared to the smaller BERT-based models.

5.2. Causal Effect of Contexts

The results for the context intervention experiments are reported in figure 4. The most desireable outcome is a DCE($S \rightarrow Y$) which is *as low as possible* in combination with a TCE(C on Y) which is *as high as possible*. For context interventions, we start to see major distinctions in the sensitivity of models to important context changes - especially the effect of the HELP fine-tuning dataset in increasing model reasoning with respect to context structure. In line with previous behavioural findings in Richardson et al. (2019a); Yanaka et al. (2019b,a); Geiger et al. (2020); Rozanova et al. (2022) and all the way back to Wang et al. (2018), which observe systematic failure of large language models in downward monotone contexts, we notice that all of the models trained only on the large benchmarks sets fail to correctly change their prediction when a context change requires it to do so (as indicated by the low TCE score). In Yanaka et al. (2019b), Rozanova et al. (2022) and Rozanova et al. (2021), the positive effect of the HELP dataset is already evident, but here we may also compare it to roberta-large-mnli tuned on many additional training sets, precluding the possibility that its helpfulness can be attributed only to a larger amount of training data.

We note that although the situation of the TCE/DCE ratio for roberta-large-mnli being less than one may seem peculiar, it is important to keep in mind that the intervention sets used for estimating these quantities are sampled independently so some margin of error is warranted. As in Stolfo et al. (2023), we interpret this result to simply mean that the causal influence is comparable whether we are affecting the ground truth result (as in the TCE(C on Y) case) or not (as in the DCE($S \rightarrow Y$) case).

5.3. Benchmark Scores and Causal Effects

A summary of the performance of all models on popular benchmarks alongside a categorical breakdown of robustness and sensitivity is presented in table 5. The robustness/sensitivity categories are a qualitative assessment, identifying the *lowest* and *highest* scores within a category, and categorising other models correspondingly as *low*, *mid* or *high* performers for the given categories. The sensitivity property is tied to the desired total causal effect, while the robustness property is tied to the undesired direct causal effect (note in particular that the latter is judged as *inversely proportional:* the model with the lowest given DCE is judged the “highest" in terms of robustness).

The key observation is that the models which

Model	NLI Benchmark Evaluation (2 Class Accuracy)						Context Changes		Inserted Word-Pair Changes	
	SNLI	MNLI-M	MNLI-MM	ANLI-R1	ANLI-R2	ANLI-R3	Robustness	Sensitivity	Robustness	Sensitivity
bert-base-uncased-snli	0.766	0.620	0.623	0.567	0.596	0.580	Mid	Mid	Mid	Low
bert-base-uncased-snli-help	0.757	0.627	0.626	0.505	0.508	0.546	Mid	Mid	Mid	Low
facebook/bart-large-mnli	0.935	0.940	0.939	0.596	0.563	0.593	High	Low	Mid	Mid
facebook/bart-large-mnli-help	0.727	0.802	0.795	0.538	0.489	0.528	Mid/High	**Highest**	**Highest**	**Highest**
roberta-large-mnli	0.931	0.941	0.940	0.614	0.529	0.5325	**Highest**	Lowest	Mid	Mid
roberta-large-mnli-help	0.738	0.668	0.656	0.565	0.554	0.574	High	**Highest**	**Highest**	**Highest**
roberta-large-snli_mnli_fever_anli	0.949	0.936	0.939	0.810	0.659	0.666	**Highest**	Lowest	Mid	Mid/High
infobert	**0.950**	**0.943**	**0.941**	**0.837**	**0.682**	**0.683**	High	Low	Mid	Mid/High

Table 5: Overall 2 class accuracy on original NLI benchmarks and qualitative comparison against the performed causal intervention analysis. The accuracy is not necessarily predictive of the performances achieved using a systematic causal inspection.

achieve the highest performance on benchmarks may be far from the best performers with respect to our quantitative markers of strong reliance of important causal features. In particular, models such as *infobert* are outperformed in our behavioural causal effect analyses by weaker models that are fine-tuned on a relatively small helper dataset such as *HELP*. It is important to note that such changes coincide with drops in benchmarks performance too, but any model interventions that discourage the exploitation of heuristics (evident from a lower DCE for surface form features) may have that effect.

6. Related Work

Natural Logic Handling in NLI Models It has been known for some time that large NLI models are frequently tripped up by downward-monotone reasoning (Richardson et al., 2019b; Wang et al., 2018; Yanaka et al., 2019b; Rozanova et al., 2022; Geiger et al., 2020). Various datasets have been created to evaluate and improve this behaviour, such as HELP (Yanaka et al., 2019b), MoNLI (Geiger et al., 2020), MQNLI (Geiger et al., 2019), MED (Yanaka et al., 2019a). Rozanova et al. (2022) introduced NLI-XY, secondary compositional dataset built from portions of MED, where the intermediate features of *context monotonicity* and *concept relations* are explicitly labelled: this is the dataset we use in this work. Non-causal structural analyses of model internals with respect to natural logic features include Rozanova et al. (2022) (a probing study), but we leave to the next section some existing works where natural logic intersects with the world of causal approaches to NLP.

Causal Analysis in NLP Causal modelling has appeared in NLP works in various forms, such as the investigations of the causal influence of data statistics (Elazar et al., 2022) and mediation analyses (Vig et al., 2020; Finlayson et al., 2021) which link intermediate linguistic/semantic features to model internals. Stolfo et al. (2023), our core reference, appears to be the first to use explicitly causal effect measures as indicators of sensitivity and robustness (for some non-causal approaches to measuring model robustness in NLP, we point to Jin et al. (2019) and Ribeiro et al. (2020)). For a fuller summary of the use of causality in NLP, please see the survey by Feder et al. (2022). Specific to natural logic, works with causal approaches include Geiger et al. (2020) (which perform interchange interventions at a token representation level), Geiger et al. (2021) (where an ambitious causal abstraction experiment attempts to align model internals with candidate causal models) and the works of Geiger et al. (2020) and Wu et al. (2022), (where attempts are made to build a prescribed causal structure into models themselves). In particular, Wu et al. (2022) create a “causal proxy model" which becomes the basis for a new explainable predictor designed to replace the original neural network.

7. Conclusion

The results strongly bolster the fact that similar benchmark accuracy scores may be observed for models that exhibit very different behaviour, especially concerning specific semantic reasoning patterns and higher-level properties such as robustness/sensitivity to target features. In this work, we have been able to causally investigate previously suspected biases in NLI models. For example, previous observations (Rozanova et al., 2022; Yanaka et al., 2019a) that roberta-large-mnli is biased in favour of assuming upward-monotone contexts, ignoring the effects of things like negation markers, agrees with our observations that it exhibits poor context sensitivity. Furthermore, the causal flavour of the study adds a complimentary narrative to works that investigate model internals via probing (Rozanova et al., 2022) and observe the presence/absence of intermediate semantic features in the models' representation. Instead of merely suggesting that these features are captured, we can gain insight into their causal influence via connected causal effect estimates. The causal measures presented here show us that even the highest-performing models can systematically fail

to adapt their predictions to changing context structure, suggesting an over-reliance on word relations across the premise and hypothesis. Finally, we have also added the observation that existing strategies to improve responsiveness to context changes also increase the *robustness* word-pair insertion changes.

Acknowledgements

This work was partially funded by the Swiss National Science Foundation (SNSF) project NeuMath (200021_204617), by the EPSRC grant EP/T026995/1 entitled "EnnCore: End-to-End Conceptual Guarding of Neural Architectures" under Security for all in an AI enabled society, by the CRUK National Biomarker Centre, and supported by the Manchester Experimental Cancer Medicine Centre.

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of NLI models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–444, St. Julian's, Malta. Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring causal effects of data statistics on language model's 'factual' predictions.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. Posing fair generalization tasks for natural language inference. 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 4485–4495, Hong Kong, China. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pages 7324–7338. PMLR.

Hai Hu and Larry Moss. 2018. Polarity computations in flexible categorial grammar. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 124–129, New Orleans, Louisiana. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*.

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

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textul inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, Prague. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019a. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019b. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Tal Linzen Richard T. McCoy. 2019. Non-entailed subsequences as a challenge for natural language inference. volume 2, pages 358–360. University of Massachusetts Amherst Libraries.

Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2019a. Probing natural language inference models through semantic fragments. In *AAAI Conference on Artificial Intelligence*.

Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2019b. Probing natural language inference models through semantic fragments. *CoRR*, abs/1909.07521.

Julia Rozanova, Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino, and Andre Freitas. 2022. Decomposing natural logic inferences for neural NLI. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 394–403, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Julia Rozanova, Deborah Ferreira, Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2021. Supporting context monotonicity abstractions in neural nli models.

Julia Rozanova, Marco Valentino, Lucas Cordeiro, and André Freitas. 2023. Interventional probing in high dimensions: An NLI case study. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2489–2500, Dubrovnik, Croatia. Association for Computational Linguistics.

Víctor Cabezas Sánchez. 1991. Studies on natural logic and categorial grammar.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. 2023. A causal framework to quantify the robustness of mathematical reasoning with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 545–561, Toronto, Canada. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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

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

of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. 2022. Causal proxy models for concept-based model explanations.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 250–255, Minneapolis, Minnesota. Association for Computational Linguistics.