# The Sensitivity of Language Models and Humans to Winograd Schema Perturbations

**Mostafa Abdou**[†] **Vinit Ravishankar**[♣]

**Maria Barrett**[†] **Yonatan Belinkov**[♠] **Desmond Elliott**[†] **Anders Søgaard**[†]

[†]Department of Computer Science, University of Copenhagen

[♣]Department of Informatics, University of Oslo

[♠]John A. Paulson School of Engineering and Applied Sciences, Harvard University

[♠]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

## Abstract

Large-scale pretrained language models are the major driving force behind recent improvements in performance on the Winograd Schema Challenge, a widely employed test of commonsense reasoning ability. We show, however, with a new diagnostic dataset, that these models are sensitive to linguistic perturbations of the Winograd examples that minimally affect human understanding. Our results highlight interesting differences between humans and language models: language models are more sensitive to number or gender alternations and synonym replacements than humans, and humans are more stable and consistent in their predictions, maintain a much higher absolute performance, and perform better on non-associative instances than associative ones. Overall, humans are correct more often than out-of-the-box models, and the models are sometimes right for the wrong reasons. Finally, we show that fine-tuning on a large, task-specific dataset can offer a solution to these issues.

## 1 Introduction

Large-scale pre-trained language models have recently led to improvements across a range of natural language understanding (NLU) tasks (Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019), but there is some scepticism that benchmark leaderboards do not represent the full picture (Kaushik and Lipton, 2018; Jumelet and Hupkes, 2018; Poliak et al., 2018). An open question is whether these models generalize beyond their training data samples.

In this paper, we examine how pre-trained language models generalize on the Winograd Schema Challenge (WSC).

Named after Terry Winograd, the WSC, in its current form, was proposed by Levesque et al. (2012) as an alternative to the Turing Test. The task takes the form of a binary reading comprehension test where a statement with two referents and a pronoun (or a possessive adjective) is given, and the correct antecedent of the pronoun must be chosen. Examples are chosen carefully to have a preferred reading, based on semantic plausibility rather than co-occurrence statistics. WSC examples come in pairs that are distinguished only by a discriminatory segment that *flips* the correct referent, as shown in Figure 1a. Levesque et al. define a set of qualifying criteria for instances and the pitfalls to be avoided when constructing examples (see §3.2). These combine to ensure an instance functions as a test of what they refer to as 'thinking' (or common sense reasoning).

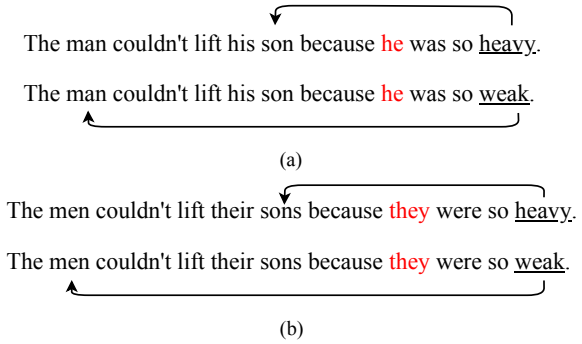


Figure 1: An example pair from the Winograd Schema Challange (a) and its perturbation (b). The pronoun resolves to one of the two referents, depending on the choice of the discriminatory segment. The perturbation in (b) pluralizes the referents and the antecedents.

Recent work has reported significant improvements on the WSC (Kocijan et al., 2019; Sakaguchi et al., 2019). As with many other NLU tasks, this improvement is primarily due to large-scale language model pre-training, followed by fine-tuning for the target task. We believe that further examination is warranted to determine whether these impressive results reflect a funda-



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7590–7604
July 5 - 10, 2020. 

mental advance in reasoning ability, or whether our models have learned to simulate this ability in ways that do not generalize. In other words, do models learn accidental correlations in our datasets, or do they extract patterns that generalize in robust ways beyond the dataset samples?

In this paper, we conduct experiments to investigate this question. We define a set of lexical and syntactic variations and perturbations for the WSC examples and use altered examples (Figure 1b) to test models that have recently reported improved results. These variations and perturbations are designed to highlight the robustness of human linguistic and reasoning abilities and to test models under these conditions.

**Contributions** We introduce a new Winograd Schema dataset for evaluating generalization across seven controlled linguistic perturbations.[1] We use this dataset to compare human and language model sensitivity to those perturbations, finding marked differences in model performance. We present a detailed analysis of the behaviour of the language models and how they are affected by the perturbations. Finally, we investigate the effect of fine-tuning with large task-specific datasets, and present an error analysis for all models.

## 2 Related Work

**Probing datasets** Previous studies have explored the robustness of ML models towards different linguistic phenomena (Belinkov and Glass, 2019), e.g., by creating challenge datasets such as the one introduced here. When predicting subject-verb agreement, Linzen et al. (2016) found that inserting a relative clause hurt the performance of recurrent networks.[2]

A large body of research has since emerged on probing pre-trained (masked) language models for linguistic structure (Goldberg, 2019; Hewitt and Manning, 2019; Lin et al., 2019; Clark et al., 2019) and analysing them via comparison to psycholinguistic and brain imaging data (Abnar et al., 2019; Ettinger, 2019; Abdou et al., 2019; Gauthier and Levy, 2019). Other recent work has attempted to probe these models for what is referred to as *common sense* or factual knowledge (Petroni et al., 2019; Feldman et al., 2019). Their findings show that these models do indeed encode such knowledge and can be used for knowledge base completion or common sense mining from Wikipedia.

**Clever Hans** A considerable amount of work has also been devoted to what might be described as the Clever Hans effect. This work has aimed to quantify the extent to which models are learning what we expect them to as opposed to leveraging statistical artifacts. This line of work has to date revealed significant problems (and some possible solutions to those problem) with reading comprehension datasets (Chen et al., 2016; Kaushik and Lipton, 2018), natural language inference datasets (Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018; Belinkov et al., 2019a; McCoy et al., 2019), and the story cloze challenge (Schwartz et al., 2017), among others.

**Winograd Schema Challenge** Trinh and Le (2018) first proposed using neural language models for the WSC, achieving an accuracy of 63.7% using an ensemble of 14 language models. Ruan et al. (2019) and Kocijan et al. (2019) fine-tune BERT (Devlin et al., 2019) on the PDP (Rahman and Ng, 2012) and an automatically generated MaskedWiki dataset, reaching an accuracy of 71.1% and 72.5% respectively. Meanwhile, Radford et al. (2019) report an accuracy of 70.7% without fine-tuning using the GPT-2 language model. Most recently, Sakaguchi et al. (2019) present an adversarial filtering algorithm which they use for crowd-sourcing a large corpus of WSC-like examples. Fine-tuning RoBERTa (Liu et al., 2019) on this, they achieve an accuracy of 90.1%.

In an orthogonal direction, Trichelair et al. (2018) presented a timely critical treatment of the WSC. They classified the dataset examples into associative and non-associative subsets, showing that the success of the LM ensemble of Trinh and Le (2018) mainly resulted from improvements on the associative subset. Moreover, they suggested switching the candidate referents (where possible) to test whether systems make predictions by reasoning about the "entirety of a schema" or by exploiting "statistical quirks of individual entities".

In a similar spirit, our work is a controlled

[1] Code and dataset can be found at: https://github.com/mhany90/enhanced_wsc/

[2] This contrasts with our results with Transformer-based architecture and is probably explained by memory loss in recurrent networks trained on short sequences. Similarly, Gulordava et al. (2018) tested whether a Recurrent Neural Network can predict long-distance number agreement in various constructions comparing natural and nonsensical sentences where RNNs cannot rely on semantic or lexical cues.

study of robustness along different axes of linguistic variation. This type of study is rarely possible in NLP due to the large size of datasets used and the focus on obtaining improved results on said datasets. Like a carefully constructed dataset which is thought to require true natural language understanding, the WSC presents an ideal testbed for this investigation.

## 3 Perturbations

We define a suite of seven perturbations that can be applied to the 285 WSC examples, which we refer to as the original examples. These perturbations are designed to test the robustness of an answer to semantic, syntactic, and lexical variation. Each of the perturbations is applied to every example in the WSC (where possible), resulting in a dataset of 2330 examples, an example of each type is shown in Table 1. Crucially, the correct referent in each of the perturbed examples is **not** altered by the perturbation. The perturbations are manually constructed, except for the sampling of names and synonyms. Further details can be found in Appendix E.

**Tense switch (TEN)** Most WSC instances are written in the past tense and thus are changed to the present continuous tense (247 examples). The remaining 34 examples are changed from the present to the past tense.

**Number switch (NUM)** Referents have their numbers altered: singular referents (and the relevant pronouns) are pluralised (223 examples), and plural referents are modified to the singular (30 examples). Sentences with names have an extra name added via conjunction; eg. "Carol" is replaced with "Carol and Susan". Possessives only mark possession on the second conjunct ("John and Steve's uncle" rather than "John's and Steve's uncle").

**Gender switch (GEN)** Each of the referents in the sentence has their gender switched by replacing their names with other randomly drawn frequent English names of the opposite gender.[3] 92% of the generated data involved a gender switch for a name. Though humans may be biased towards gender (Collins, 2011; Desmond and Danilewicz, 2010; Hoyle et al., 2019), the perturbations do not introduce ambiguity concerning gender, only the entity. 101 examples were switched from male to female, and 55 examples the other way around.

**Voice switch (VC)** All WSC examples, except for 210 and 211, are originally in the active voice and are therefore passivized. 210 and 211 are changed to the active voice. 65 examples could not be changed. Passive voice is known to be more difficult to process for humans (Olson and Filby, 1972; Feng et al., 2015).

**Relative clause insertion (RC)** A relative clause is inserted after the first referent. For each example, an appropriate clause was constructed by first choosing a template such as "who we had discussed" or "that is known for" from a preselected set of 19 such templates. An appropriate ending, such as "who we had discussed *with the politicians*" is then appended to the template depending on the semantics of the particular instance. Relative clauses impose an increased demand on working memory capacity, thereby making processing more difficult for humans (Just and Carpenter, 1992; Gibson, 1998).

**Adverbial qualification (ADV)** An adverb is inserted to qualify the main verb of each instance. When a conjunction is present both verbs are modified. For instances with multiple sentences, all main verbs are modified.

**Synonym/Name substitution (SYN/NA)** Each of the two referents in an example is substituted with an appropriate synonym, or if it is a name, is replaced with a random name of the same gender from the same list of names used for the gender perturbation.

### 3.1 Human Judgments

We expect that humans are robust to these perturbations because they represent naturally occurring phenomena in language; we test this hypothesis by collecting human judgements for the perturbed examples. We collect the judgments for the perturbed examples using Amazon Mechanical Turk. The annotators are presented with each instance where the pronoun of interest is boldfaced and in red font. They are also presented with two options, one for each of the possible referents. They are then instructed to choose the most likely option, in exchange for $0.12. Following Sakaguchi et al. (2019), each instance is annotated by three anno-

[3] Names sourced from https://github.com/AlessandroMinoccheri/human-names/tree/master/data

| | Instance / Perturbed Instance | Count |
|---|---|---|
| Original | Sid explained his theory to Mark but he couldn't convince him. | 285 |
| Tense | Sid is explaining his theory to Mark but he can't convince him. | 281 |
| Number | **Sid and Johnny** explained their theory to **Mark and Andrew** but they couldn't convince them. | 253 |
| Gender | **Lucy** explained her theory to **Emma** but she couldn't convince her. | 155 |
| Voice | The theory was explained by Sid to Mark but he couldn't convince him. | 220 |
| Relative clause | Sid, **which we had seen on the discussion panel with Chris,** explained his theory to Mark but he couldn't convince him. | 283 |
| Adverb | Sid **diligently** explained his theory to Mark but he couldn't convince him. | 283 |
| Synonyms/Names | **John** explained his theory to **Jad** but he couldn't convince him. | 285 |

Table 1: Examples from our dataset of the different perturbations applied to a WSC instance.

tators and majority vote results are reported. Results are reported later in §5. All three annotators agreed on the most likely option in 82-83% of the instances, except for gender, where a full agreement was obtained for only 78% of the instances. See Appendix B for further annotation statistics, a sample of the template presented to annotators, and restrictions applied to pool of annotators. We did not require an initial qualification task to select participants.

### 3.2 Confounds and Pitfalls

Constructing WSC problems is known to be difficult. Indeed, the original dataset was carefully crafted by domain experts and subsequent attempts at creating WSC-like datasets by non-experts such as in Rahman and Ng (2012) have produced examples which were found to be less challenging than the original dataset. Two likely pitfalls listed in Levesque et al. (2012) concern **A**) statistical preferences which make one answer more readily associated with the special discriminatory segment or other components of an example[4] (this is termed as *Associativity*, and it is described as *non-Google-proofness* in Levesque et al. (2012)); and **B**) inherent ambiguity which makes the examples open to other plausible interpretations. In what follows, we discuss these pitfalls, demonstrating that the perturbed examples remain resilient to both.

**Quantifying Associativity** To verify that the perturbations have not affected the correctness of the original problems with regards to pitfall **A**, we employ pointwise mutual information (PMI) to test the associativity of both the original and perturbed examples. PMI is known to be a reasonable measure of associativity (Church and Hanks, 1990) and, among a variety of measures, has been shown to correlate best with association scores from human judgements of contextual word association (Frassinelli, 2015). We compute unigram PMI on the two corpora used to train BERT (see Appendix C for details). Figure 2 shows the *divergence* of the perturbed examples from the original WSC dataset. We estimate divergence as the average difference in PMI between the correct ($\mathcal{C}$) and incorrect ($\mathcal{I}$) candidates: $\Delta = pmi(c_j, x_j) - pmi(i_j, x_j)$ where $\mathcal{X}$ is either: i) the discriminatory segments or ii) the full text of the example, and $pmi(\cdot, \cdot)$ is average unigram PMI. $\Delta$ can be

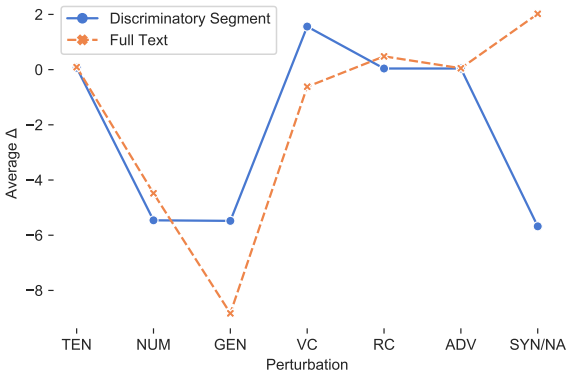


Figure 2: PMI divergence from the original WSC examples in average $\Delta$ for each perturbation. Values below 0 indicate that the difference in PMI between the correct candidate and the incorrect one decreased.

[4] Trichelair et al. (2018) find that 13.5% of examples from the original WSC might still be considered to be *associative*.

seen as a measure of whether the correct or incorrect candidate is a better 'associative fit' for either the discriminatory segment or the full context, making the examples trivial to resolve. Observe that this difference in PMI declines for the perturbed examples, showing that these the perturbed example do not increase in associativity.

**Confirming Solvability** Three expert annotators[5] are asked to solve the small subset of examples (99 in total across perturbations) which were annotated incorrectly by the majority vote of Mechanical Turk workers. To address pitfall **B**, the expert annotators are asked to both attempt to solve the instances and indicate if they believe them to be *too ambiguous* to be solved. The majority vote of the annotators determines the preferred referent or whether an instance is ambiguous. Out of a total of 99 examples, 10 were found to be ambiguous. Of the remaining 89 examples, 67 were answered correctly by the majority vote. See Appendix D for more details.

## 4 Experimental Protocol

Our experiments are designed to test the robustness of language models to the Winograd Schema perturbations described in the previous section.

**Evaluation** Models are evaluated using two types of measures. The first is *accuracy*. For each of the perturbations, we report (a) the accuracy on the perturbed set (**Perturbation accuracy**), (b) the difference in accuracy on the perturbed set and on the *equivalent subset* of original dataset:[6] $\Delta_{\textbf{Acc.}}$ = **Perturbation accuracy** − **Original subset accuracy**, and (c) **Pair accuracy**, defined as the number of pairs for which both examples in the pair are correctly answered divided by the total number of pairs.

The second measure is *stability*, $S$. This is the proportion of perturbed examples $\mathcal{P}'$ for which the predicted referent is the same as the original prediction $\mathcal{P}$:

$$S = \frac{\mid \{(p'_i, p_i) \mid p'_i \in \mathcal{P}' \wedge p_i \in \mathcal{P} \wedge p'_i = p_i\} \mid}{\mid \mathcal{P} \mid}$$

Since the perturbations do not alter the correct referent, this provides a strong indication of robustness towards them.

[5] Graduate students of linguistics.

[6] Recall that is was not possible to perturb all examples.

**Baseline** We take the unigram PMI between candidates and discriminatory segments (see §3.2) as a baseline. We expect that this simple baseline will perform well for instances with a high level of associativity but not otherwise.

**Language Models** Our analysis is applied to three out-of-the-box language models (LMs): BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019). These models are considered to be the state-of-the-art for the wide variety of natural language understanding tasks found in the GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks. We use the *large* pre-trained publicly available models (Wolf et al., 2019).[7]

**Fine-tuned Language Models** We also examine the effect of fine-tuning language models. BERT+WW uses BERT fine-tuned on the MaskedWiki and WscR datasets which consist of 2.4M and 1322 examples (Kocijan et al., 2019), and RoBERTa+WG is fine-tuned on WinoGrande **XL**, which consists of 40,938 adversarially filtered examples (Sakaguchi et al., 2019). Both fine-tuned models have been reported by recent work to achieve significant improvements on the WSC.

**Scoring** To score the two candidate referents in each WSC instance we employ one of two mechanisms. The first, proposed in Trinh and Le (2018) and adapted to masked LMs by Kocijan et al. (2019) involves computing the probability of the two candidates $c1$ and $c2$, given the rest of the text in the instance $s$. To accomplish this, the pronoun of interest is replaced with a number of MASK tokens corresponding to the number of tokens in each of $c1$ and $c2$. The probability of a candidate, $p(c|s)$ is then computed as the average of the probabilities assigned by the model to the candidate's tokens and the maximum probability candidate is taken as the answer. This scoring method is used for all models, except RoBERTa+WG. For that, we follow the scoring strategy employed in Sakaguchi et al. (2019) where an instance is split into context and option using the candidate answer as a delimiter.[8]

[7] https://github.com/huggingface/pytorch-transformers

[8] [CLS] context [SEP] option [SEP], e.g. *[CLS] The sculpture rolled off the shelf because ____ [SEP] wasn't anchored [SEP].* The blank is filled with either option 1 (*the sculpture*) or 2 (*the trophy*).

## 5 Results and Analysis

Following the experimental protocol, we evaluate the three out-of-the-box language models and the two fine-tuned models on the original WSC and each of the perturbed sets. Table 2 shows **Perturbation accuracy** results for all models[9] and contrasts them with human judgements and the PMI baseline.

### 5.1 Language Models

Humans maintain a much higher performance compared to out-of-the-box LMs across perturbations. The difference in accuracy between the perturbed and original examples, $\Delta_{\textbf{Acc.}}$, as defined in Section 4 is shown in Figure 4. A general trend of decrease can be observed for both models and humans across the perturbations. This decline in accuracy is on average comparable between models and humans — with a handful of exceptions. Taking the large gap in absolute accuracy into account, this result might be interpreted in two ways. If a comparison is made relative to the upper bound of performance, human performance has suffered from a larger error increase. Alternately, if we compare relative to the lower bound of performance, then the decline in the already low performance of language models is more meaningful, since 'there is not much more to lose'.

A more transparent view can be gleaned from the stability results shown in Table 3. Here it can be seen that the three out-of-the-box LMs are *substantially more likely* to switch predictions due to the perturbations than humans. Furthermore, we observe that the LMs are least stable for word-level perturbations like gender (GEN), number (NUM), and synonym or name replacement (SYN/NA), while humans appear to be most affected by sentence-level ones, such as relative clause insertion (RC) and voice perturbation (VC).

**Understanding Language Model Performance**

To better understand the biases acquired through pre-training which are pertinent to this task, we consider a) a case of essential feature omission and b) the marginal cases where LMs answer very correctly or incorrectly, in both the original and perturbed datasets. We present analysis for BERT, but similar findings hold for the other LMs.

**Masking discriminatory segments** result in identical sentence pairs because these segments are the only part of a sentence that sets WSC pairs apart (see Figure 1a). To determine whether there is a bias in the selectional preference for one of the candidates over the other, we test BERT on examples where these discriminatory segments have been replaced with the `MASK` token. An unbiased model should be close to random selection but BERT consistently prefers (by a margin of ∼25-30%) the candidate which appears second in the text to the one appearing first, for all perturbations except voice, where it prefers the first. This observation holds even when the two referents are inverted, which is possible for the 'switchable' subset of the examples as shown in Trichelair et al. (2018). This indicates that the selections are not purely semantic but also syntactic or structural and it points towards BERT having a preference referents in the object role. Detailed results are presented in Appendix F.

**Marginal examples** are found where the model assigns a much higher probability to one referent over the other. We extract the top 15% examples where the correct candidate is preferred by the largest margin ($\mathrm{P}_{\texttt{correct}} \gg \mathrm{P}_{\texttt{incorrect}}$) and the bottom 15% where the incorrect one is preferred ($\mathrm{P}_{\texttt{incorrect}} \gg \mathrm{P}_{\texttt{correct}}$). Surprisingly, we find that there is a large overlap (50%–60%) between these two sets of examples, both in the original and the perturbed datasets.[10] For the examples which are both the most correct and incorrect, BERT strongly prefers one of the candidates without considering the special discriminatory segment which *flips* the correct referent. Indeed we find that the correlation between the probability assigned by BERT to a referent when it is the correct referent and when it is not is very strong and significant, with Spearman's $\rho \approx 0.75$ across perturbations (see Appendix G for details).

[9] It is interesting to note that XLNet is trained on CommonCrawl which indexes an online version of the original WSC found here: `https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html`.

[10] To clarify, consider the following original WSC pair:

(i) Alice looked for her friend Jade in the crowd. Since she always has good luck, Alice spotted her quickly.

(ii) Alice looked for her friend Jade in the crowd. Since she always wears a red turban, Alice spotted her quickly.

The first example gives $\mathrm{P}_{\texttt{correct}} \gg \mathrm{P}_{\texttt{incorrect}}$ by the largest margin, and its counterpart gives $\mathrm{P}_{\texttt{incorrect}} \gg \mathrm{P}_{\texttt{correct}}$ by the largest margin. In other words, the model assigns a *much higher probability* for *Alice* in both cases.

| | ORIG | TEN | NUM | GEN | VC | RC | ADV | SYN/NA | *Avg* | *Avg* $\Delta_{\textbf{Acc.}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| PMI | 54.38 | 54.09 | 52.96 | 57.42 | 54.09 | 54.41 | 54.41 | 51.92 | 54.24 | −2.13 |
| BERT | 61.75 | 61.92 | 57.31 | 57.42 | 63.64 | 62.19 | 61.48 | 58.59 | 60.41 | −1.26 |
| XLNET | 64.56 | 60.14 | 62.45 | 62.58 | 57.73 | 62.9 | 64.31 | 61.05 | 61.59 | −2.78 |
| ROBERTA | 69.82 | 69.40 | 64.43 | 53.55 | 66.82 | 68.55 | 69.61 | 57.54 | 64.27 | −5.16 |
| BERT+WW | 72.28 | 70.46 | 71.15 | 74.84 | 65.91 | 64.31 | 72.44 | 70.88 | 70.00 | −2.82 |
| ROBERTA+WG | 88.42 | 89.32 | 88.53 | 86.45 | 83.63 | 86.93 | 88.7 | 89.05 | 87.62 | −1.06 |
| HUMANS | 97.89 | 96.79 | 94.46 | 92.25 | 92.27 | 91.16 | 95.40 | 96.14 | 94.41 | −3.83 |

Table 2: Original dataset accuracy (ORIG) and **Perturbation accuracy** results for all models and humans. The penultimate column shows the average **Perturbation accuracy** results. The rightmost column shows the $\Delta_{\textbf{Acc.}}$ results, averaged over all perturbations.

| | TEN | NUM | GEN | VC | RC | ADV | SYN/NA | *Avg* |
|---|---|---|---|---|---|---|---|---|
| PMI | 100 | 100 | 73.91 | 100 | 100 | 100 | 100 | 96.27 |
| BERT | 89.32 | 69.17 | 88.39 | 79.55 | 83.75 | 91.87 | 68.42 | 81.40 |
| XLNET | 82.21 | 69.17 | 66.45 | 69.55 | 78.45 | 84.81 | 70.53 | 75.02 |
| ROBERTA | 91.46 | 77.47 | 61.29 | 79.09 | 83.75 | 89.75 | 68.77 | 79.26 |
| BERT+WW | 90.04 | 83.00 | 89.68 | 80.45 | 81.98 | 92.93 | 85.96 | 85.14 |
| ROBERTA+WG | 96.08 | 94.07 | 97.41 | 91.36 | 92.22 | 94.69 | 96.11 | 95.24 |
| HUMANS | 96.70 | 94.9 | 92.9 | 91.18 | 91.11 | 96.11 | 96.1 | 94.31 |

Table 3: Stability results for all models and humans.

### 5.2 The effect of fine-tuning

The accuracy and stability results (Tables 2 and 3) indicate that fine-tuning makes language models more robust to the perturbations. ROBERTA+WG, in particular, is the most accurate and most stable model. While impressive, this is not entirely surprising: fine-tuning on task-specific datasets is a well-tested recipe for bias correction (Belinkov et al., 2019b). Indeed, these results provide evidence that it is possible to construct larger fine-tuning datasets whose distribution is correct for the WSC. We note that both fine-tuned models perform worst on the VC and RC perturbations, which may not frequently occur in the crowd-sourced datasets used for fine-tuning. To test this intuition, we apply a dependency parser (UDPipe (Straka et al., 2016)) to the WinoGrande XL examples, finding that only $\sim 5\%$ of the examples are in the passive voice and $\sim 6.5\%$ contain relative clauses.

**How much fine-tuning data is needed?** To quantify the amount of fine-tuning data needed to achieve robustness, we fine-tune ROBERTA on the five WinoGrande training set splits defined by Sakaguchi et al. (2019): **XS** (160)[11], **S** (640), **M** (2558), **L** (10234), and **XL** (40398). Figure 3 shows the average accuracy and stability scores for the models fine-tuned on each of the training splits[12]. We observe that the two smallest splits do not have a sufficient number of examples to adequately bias the classification head, leading to near-random performance. The model fine-tuned on the **M** split—with just 2558 examples—is, however, already able to vastly outperform the non-fine-tuned ROBERTA. Increasing the number of examples five-fold and twenty-fold leads to significant but fast diminishing improvements.

[11] No. of examples in set.

[12] Note that the stability score for the model fine-tuned on **XL** in Figure 3 is different from that reported in Table 3. In the latter we reported results from the model provided by Sakaguchi et al. (2019), rather than the model we fine-tuned ourselves. Since we utilise identical hyperparameters to theirs for fine-tuning, this anomalous difference in score may perhaps be explained by a difference in initialization as suggested in Dodge et al. (2020).

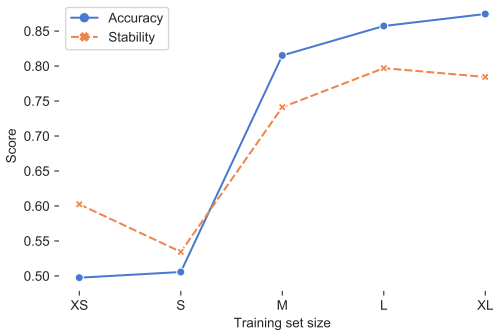


Figure 3: Accuracy and stability scores (averaged across perturbations) for RoBERTA when fine-tuned on five increasing training split sizes.

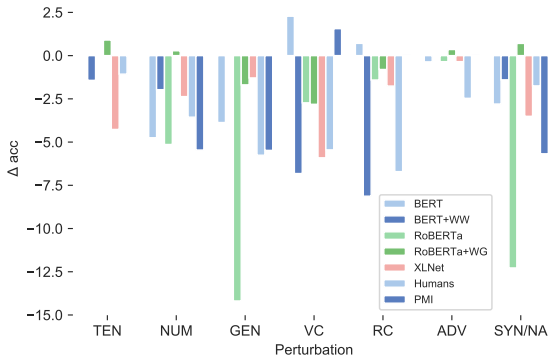


Figure 4: $\Delta_{\text{Acc.}}$ results for all models across perturbations. Values below the x-axis indicate a decline in accuracy compared to the original dataset.

**How do perturbations affect token probability distributions?** To obtain a holistic view of the effect the perturbations have on LMs and fine-tuned LMs, we analyze of the shift in the probability distribution (over the entire vocabulary) which a model assigns to a `MASK` token inserted in place of the pronoun of interest. We apply probability distribution truncation with a threshold of $p = 0.9$ as proposed in Holtzman et al. (2019) to filter out the uninformative tail of the distribution. Following this, we compute the Jensen–Shannon distance between this dynamically truncated distribution for an original example and each of its perturbed counterparts. Figure 5 shows the average of this measure over the subset of the 128 examples which are common to all perturbations. Overall, we observe that large shifts in the distribution correspond to lower stability and accuracy scores and that fine-tuned models exhibit lower shifts than their non-fine-tuned counterparts. The difference in shifts between out-of-the-box models and their fine-tuned counterparts is lower for the VC, RC and ADV perturbations, meaning that when fine-tuned, the models' probability distributions are roughly just as divergent for these perturbations as they were before fine-tuning. We hypothesize the same reasons we did in 5.2, which is that these examples are just under-represented in our fine-tuning corpus; indeed, these results roughly correspond to the differences in $\Delta_{\text{Acc.}}$ from Figure 4.

Further details about the number of examples excluded via the probability distribution truncation and other measures of the perturbations' effect can be found in Appendix G.

### 5.3 Error Analysis

**Pair Accuracy** Here we consider a more challenging evaluation setting where each WSC pair is treated as a single instance. Since the WSC examples are constructed as minimally contrastive pairs (Levesque et al., 2012), we argue that this is an appropriate standard of evaluation. Consider again the example in Figure 1a. It is reasonable to suppose that for an answerer which truly 'understands' (Levesque et al., 2012), being able to link the concepts *heavy* and *son* in one of the resolutions is closely related and complementary to linking the concepts *weak* and *man* in the other.[13]

The results for this evaluation are shown in Figure 6. They show that human resolution of the problems exhibits greater complementarity compared to the language models; human pair accuracy (pair) is closer to perturbation accuracy (single) than is the case for the LMs. Furthermore, human performance on pair accuracy is more robust to perturbations when compared to the models. Indeed, the large gap between pair accuracy and perturbation accuracy raises some doubts about the performance of these models. However, RoBERTa-WG is a notable exception, showing near-human robustness to pair complementarity.

**Associativity** Next, we examine the effect of associativity on performance. Figure 7 shows accuracy results[14] for all perturbations on the associative and non-associative subsets of the WSC as labelled by Trichelair et al. (2018). We observe that the difference between associative and non-

[13] As a sanity check, consider random pairings of WSC examples. There is no such complement.

[14] Note that the large variance in results on the associative subset of gender is due to it consisting of only two examples.

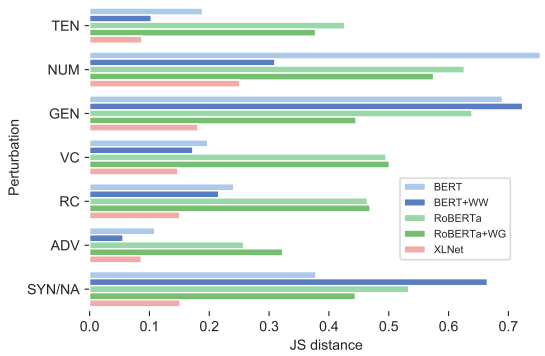


Figure 5: Jensen-Shannon distance between the original and perturbed examples when masking the pronoun of interest.

associative is much smaller for humans and that unlike all language models, humans do better on the former than the latter. As expected, the PMI baseline does almost as well as the LMs on the associative subset but it performs at chance level for the non-associative subset.

## 6 Conclusion

We presented a detailed investigation of the effect of linguistic perturbations on how language models and humans perform on the Winograd Schema Challenge. We found that compared to out-of-the-box models, humans are significantly more stable to the perturbations and that they answer non-associative examples with higher accuracy than associative ones, show sensitivity to WSC pair complementarity, and are more sensitive to sentence-level (as opposed to word-level) perturbations. In an analysis of the behaviour of language models, we observe that there is a preference for referents in the object role and that the models do not always consider the discriminatory segments of examples. Finally, we find that fine-tuning language models can lead to much-improved accuracy and stability. It remains an open question whether this task-specific approach to generalisation constitutes a true advancement in "reasoning". Fine-tuning a model on a rather large number of examples similar to the WSC leads to increased robustness, but this stands in stark contrast to humans, who are robust to the perturbations without having been exposed to similar examples in the past.

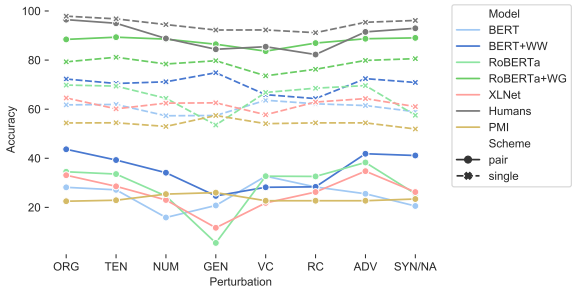


Figure 6: **Pair accuracy** and **Perturbation accuracy** results. The latter are labeled as *single*.

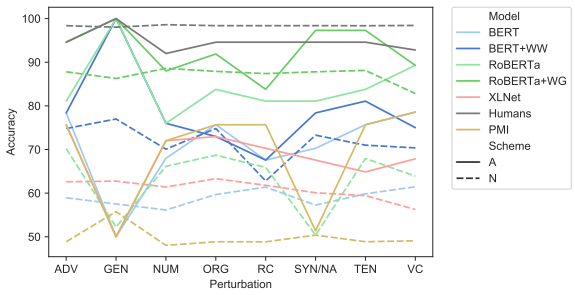


Figure 7: Perturbation accuracy on the Associative (A) and Non-Associative (N) subsets of the data.

## Acknowledgments

We would like to thank Mitja Nikolaus, Artur Kulmizev, Ana Valeria Gonzalez, and the anonymous reviewers for their helpful comments. Mostafa Abdou and Anders Søgaard are supported by a Google Focused Research Award and a Facebook Research Award. Yonatan Belinkov was supported by the Harvard Mind, Brain, and Behavior Initiative.

## References

Mostafa Abdou, Artur Kulmizev, Felix Hill, Daniel M Low, and Anders Søgaard. 2019. Higher-order comparisons of sentence encoder representations. *arXiv preprint arXiv:1909.00303*.

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. *arXiv preprint arXiv:1906.01539*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019a. On

adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 256–262, Minneapolis, Minnesota. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart M Shieber, Benjamin Van Durme, and Alexander M Rush. 2019b. Don't take the premise for granted: Mitigating artifacts in natural language inference. *arXiv preprint arXiv:1907.04380*.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Rebecca L Collins. 2011. Content analysis of gender roles in media: Where are we now and where should we go? *Sex roles*, 64(3-4):290–298.

Roger Desmond and Anna Danilewicz. 2010. Women are on, but not in, the news: Gender roles in local television news. *Sex Roles*, 62(11-12):822–829.

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

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Allyson Ettinger. 2019. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *arXiv preprint arXiv:1907.13528*.

Joshua Feldman, Joe Davison, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained models.

Shiwen Feng, Jennifer Legault, Long Yang, Junwei Zhu, Keqing Shao, and Yiming Yang. 2015. Differences in grammatical processing strategies for active and passive sentences: An fmri study. *Journal of Neurolinguistics*, 33:104–117.

Diego Frassinelli. 2015. The effect of context on the activation and processing of word meaning over time.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. *arXiv preprint arXiv:1910.01244*.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt 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.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Marcel A Just and Patricia A Carpenter. 1992. A capacity theory of comprehension: individual differences in working memory. *Psychological review*, 99(1):122.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside bert's linguistic knowledge. *arXiv preprint arXiv:1906.01698*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

David R Olson and Nikola Filby. 1972. On the comprehension of active and passive sentences. *Cognitive Psychology*, 3(3):361–381.

Fabio Petroni, Tim Rocktschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789. Association for Computational Linguistics.

Yu-Ping Ruan, Xiaodan Zhu, Zhen-Hua Ling, Zhan Shi, Quan Liu, and Si Wei. 2019. Exploring unsupervised pretraining and sentence structure modelling for winograd schema challenge. *arXiv preprint arXiv:1904.09705*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. *arXiv preprint arXiv:1702.01841*.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Paul Trichelair, Ali Emami, Jackie Chi Kit Cheung, Adam Trischler, Kaheer Suleman, and Fernando Diaz. 2018. On the evaluation of common-sense reasoning in natural language understanding. *arXiv preprint arXiv:1811.01778*.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. *arXiv preprint arXiv:1804.08117*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

| Pert. | Full Agreement | Avg. Time |
|---|---|---|
| ORG | 82.45 | 15.32 |
| TEN | 82.91 | 16.39 |
| NUM | 83.00 | 19.56 |
| GEN | 78.06 | 19.24 |
| VC | 82.72 | 17.02 |
| RC | 82.68 | 17.83 |
| ADV | 82.68 | 17.69 |
| SYN/NA | 82.45 | 15.26 |

Table 4: Annotation statistics: Proportion of examples with full agreement and average time required for answering in seconds.

## A Observations on original dataset

1. A few of the original examples were of unorthodox design: for instance, consider the pair:

   (1) a. Look! There is a minnow swimming right below that duck! It had better get away to safety fast!

   b. Look! There is a shark swimming right below that duck! It had better get away to safety fast!

   Here, instead of having a discriminatory segment select which of the two nouns could be the antecedent, one of the nouns is switched out with another.

2. Example 90 has a typo in the question where Kamchatka is spelled as 'Kamtchatka'.

## B Human Judgements

Table 4 shows the proportion of instances for which all three annotators agreed and the average time required by annotators for the original examples and each of the perturbed datasets. Figure 8 shows the Amazon Mechanical Turk template used. The annotator pool was restricted to native speakers of English located in the United States who were classified by Mturk as 'masters' and had a HITs approval rate above 99%.

## C Pointwise Mutual Information

We compute unigram Pointwise Mutual Information statistics using the Hyperwords[15] package (Levy et al., 2015). If a corpus is split into a collection $D$ of words $W$ and their contexts $C$, we

[15] https://bitbucket.org/omerlevy/hyperwords/

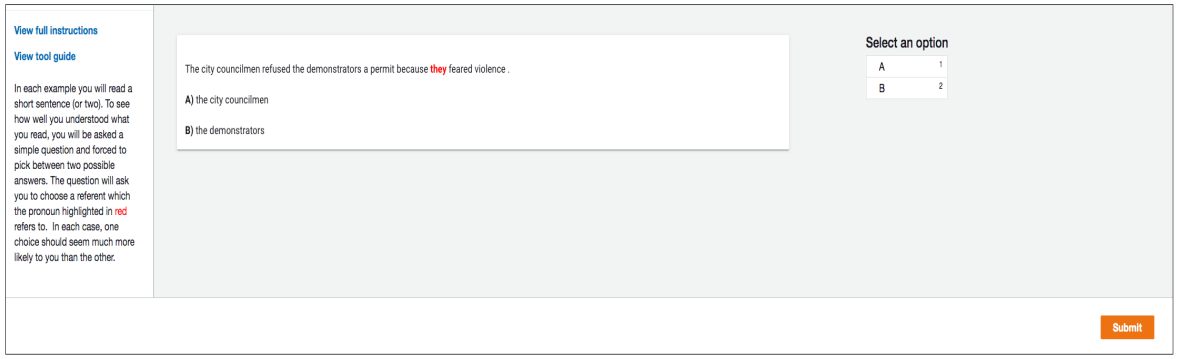


Figure 8: Sample of Mturk template shown to annotators.

can compute co-occurrence counts for each pair of $w \in W$ and $c \in C$. PMI is then defined as the log-ratio between the joint probability of $w$ with $c$ and the product of their marginal probabilities. Refer to Levy et al. (2015) for further details. For generating a collection $D$ of word-context pairs, we use the following hyperparameter settings: a minimal word count of 200 for being in the vocabulary, a context window size of 6, dynamic context windows, positional contexts (where each context is a conjunction of a word and its relative position to the target word).

## D Confirming Solvability

Table 5 shows the breakdown by perturbation type of the expert annotations which were gathered for examples that were annotated incorrectly by the Mechanical Turk workers.

## E Notes on construction of perturbed dataset

**Tense switch (TEN)** Examples 168–172 could not be changed while maintaining the semantics of the instance intact.

**Relative clause insertion (RC)** The pre-selected set of 19 templates is shown below:

| Counts | All | Ambig. | Non-Ambig. | Correct |
|---|---|---|---|---|
| TEN | 9 | 0 | 9 | 8 |
| NUM | 14 | 2 | 12 | 9 |
| GEN | 12 | 2 | 10 | 10 |
| VC | 17 | 3 | 14 | 12 |
| RC | 25 | 1 | 24 | 13 |
| ADV | 13 | 0 | 13 | 11 |
| SYN/NA | 9 | 2 | 7 | 4 |

Table 5: Breakdown of solvability annotation counts by perturbation. **Ambig.** indicates the count of examples labeled as Ambiguous, **Non-Ambig.** is the number of remaining examples. **Correct** indicates the number of those which is solved correctly.

- "who we had discussed __"
- "who he had discussed __"
- "who she had discussed __"
- "who you had discussed __"
- "which we had seen __"
- "which he had seen __"
- "which she had seen __"
- "which you had seen __"
- "who we know from __"
- "who he knows from __"
- "who she knows from __"
- "who you know from __"
- "that is mentioned in __"
- "that is located at __"
- "that is close to __"
- "that is known for __"
- "which had been __",
- "who you met __"
- "that is __"
- "which was put there __"

**Synonym/Name substitution (SYN/NA)** No appropriate synonyms were found for *tide* and *wind* in examples 130 and 131.

**Adverbial qualification (ADV)** Two instances (95 and 96) in which the main verb was already modified were excluded.

## F Referent preferences

Table 6 shows the percentage of examples in the switchable subset of the datasets where the second referent in the text was assigned a higher probability than the first, for both the original and reversed referent order.

## G Effect of perturbations

**Nucleus Sampling** Table 7 shows the average number of vocabulary items kept after Nucleus sampling with $p = 0.9$ is applied.

| Pert. | Original | Reversed |
|---|---|---|
| ORG | 66.90 | 70.42 |
| TEN | 62.38 | 65.14 |
| NUM | 60.16 | 56.10 |
| GEN | 72.17 | 75.65 |
| VC | 38.14 | 39.83 |
| RC | 63.57 | 68.57 |
| ADV | 68.08 | 70.92 |
| SYN/NA | 59.12 | 64.23 |

Table 6: Percentage of examples in switchable subset with probabilities assigned to the second referent in the text rather than the first, for both the original and reversed referent order.

**Probability shift** is defined as the difference in the probability of a candidate before and after a perturbation is applied. Figure 9 shows the difference in average probability shift between the correct candidates and the incorrect candidates for each of the models per perturbation type. This provides a view that is meaningfully different from accuracy, as the probability of a candidate can shift without exceeding the threshold required to change a model's prediction. We find that there is a general trend of the incorrect candidates becoming more likely relative to the correct ones. This can be seen as confirming that, on average, nearly all perturbations make the problems more difficult for all models.

**Hidden state representation distance** is used to provide a more holistic view of the correspondence between the representations derived for the different perturbations. The analysis is conducted on the 128 examples which are common between all datasets. A representation is derived for each example by taking the max-pool of hidden-state representations of a model's final layer. For each of the seven perturbations $p$, we compute pairwise correlation distance[16] between each pair of original and perturbed example representations yielding a vector $\vec{D_p} \in \mathbb{R}^{128}$. The mean of $\vec{D_p}$ is then computed as an aggregate measure of the distance between the representations derived from a perturbation $p$ and the original $o$. Figure 10 shows a plot of this for all perturbations for each of the models.

[16]This is preferable to other distance measures as it normalizes both the mean and variance of activity patterns over experimental conditions.

| Perturbation | BERT | RoBERTa | XLNet | BERT+WW | RoBERTa+WG |
|---|---|---|---|---|---|
| ORG | 19.81 | 203 | 1.26 | 1.07 | 1021.44 |
| TEN | 23.88 | 165.84 | 1.26 | 1.09 | 947.53 |
| NUM | 90.35 | 341.05 | 1.57 | 1.30 | 1087.78 |
| GEN | 18.11 | 128.37 | 1.44 | 1.19 | 1039.84 |
| VC | 41.88 | 154.21 | 1.28 | 1.09 | 961.04 |
| RC | 21.02 | 97.35 | 1.35 | 1.14 | 952.09 |
| ADV | 17.01 | 145.35 | 1.23 | 1.10 | 1004.14 |
| SYN/NA | 31.50 | 199.26 | 1.39 | 1.11 | 1055.71 |
| VOCAB. SIZE | 30522 | 50265 | 32000 | 30522 | 50265 |

Table 7: Average number of vocabulary items left after probability distribution truncation with $p = 0.9$ is applied.

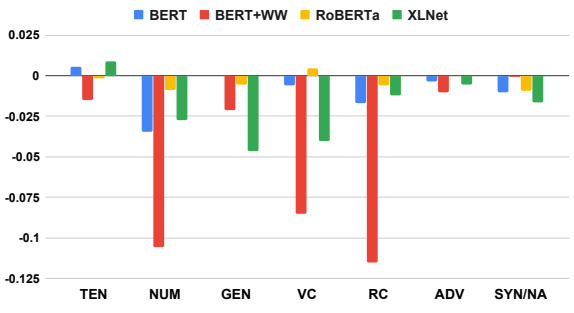


Figure 9: The difference between average probability shift for the correct and the incorrect referents per perturbation. Y-axis values above zero mean the correct referent became more likely on average after a perturbation and vice versa.

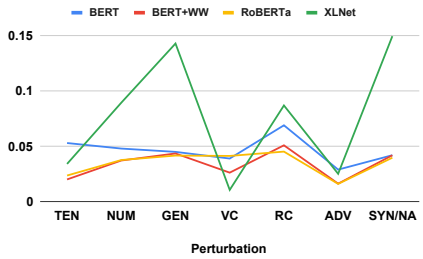


Figure 10: The correlation of pronoun hidden state representation distance from the original for each perturbation.

## H Candidate probability correlations

Figure 11 shows the average correlation between a candidate's probability when it is the correct referent and when it is not.

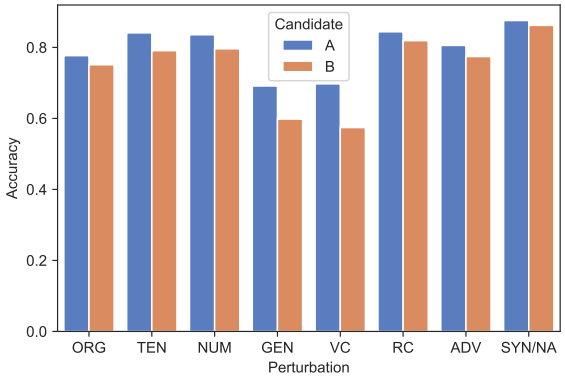


Figure 11: Correlation (Spearman's $\rho$) between the probability of a candidate when it is the correct candidate and when it is the incorrect one. Candidates A and B are the first and second candidates in a WSC instance.