

Back to Square One: Bias Detection, Training and Commonsense Disentanglement in the Winograd Schema

Yanai Elazar^{1,2} Hongming Zhang^{3,4} Yoav Goldberg^{1,2} Dan Roth⁴

¹Bar Ilan University, ²AI2, ³HKUST, ⁴UPenn

{yanaiel, yoav.goldberg}@gmail.com

hzhangal@cse.ust.hk, danroth@seas.upenn.edu

Abstract

The Winograd Schema (WS) has been proposed as a test for measuring commonsense capabilities of models. Recently, pre-trained language model-based approaches have boosted performance on some WS benchmarks but the source of improvement is still not clear. We begin by showing that the current evaluation method of WS is sub-optimal and propose a modification that makes use of twin sentences for evaluation. We also propose two new baselines that indicate the existence of biases in WS benchmarks. Finally, we propose a method for evaluating WS-like sentences in a zero-shot setting and observe that popular language models perform randomly in this setting. We conclude that much of the apparent progress on WS may not necessarily reflect progress in commonsense reasoning, but much of it comes from supervised data, which is not likely to account for all the required commonsense reasoning skills and knowledge.

1 Introduction

The Winograd Schema (WS) (Levesque et al., 2012) was proposed as an alternative to the Turing test, by virtue of evaluating progress on commonsense reasoning. The task is a multi-choice question akin to coreference resolution. Given a text snippet with two entities and a pronoun that refers to one of the entities, select the entity referred to by the pronoun.¹ Consider the following example:

1. *The trophy doesn’t fit in the brown suitcase because **it** was too large.*

The entities are marked in italics, the pronoun in bold, and the special word² is underlined. In this

¹It can also be a possessive adjective, but for simplicity, we refer these as pronouns.

²Words that change the answer. A detailed explanation is provided later.

Example	Answer	Setup
<i>The trophy doesn’t fit in the brown suitcase because it was too <u>large</u>.</i>	<u>trophy</u>	<u>twin-1</u>
<i>The trophy doesn’t fit in the brown suitcase because it was too <u>small</u>.</i>	<u>suitcase</u>	<u>twin-2</u>
<i>doesn’t fit in because it was too <u>large</u>. because it was too <u>large</u>.</i>	?	<u>Baselines</u> <i>no-cands</i> <i>part-sent</i>
<i>The trophy doesn’t fit in the brown suitcase because the trophy was too [MASK].</i>	<u>large</u>	<u>Zero-shot</u> <u>twin-1</u>
<i>The trophy doesn’t fit in the brown suitcase because the brown suitcase was too [MASK].</i>	<u>small</u>	<u>twin-2</u>

Figure 1: Examples from the Winograd Schema Challenge (top), our proposed modification to these sentences that we use as novel baselines (middle) and the new formulation of the WS task which allows us to test LMs in a zero-shot setting (bottom).

case, **it** refers to *The trophy*, since smaller objects typically fit into larger objects.³

The success of Pretrained Language Models (PLMs) seems to have advanced models’ commonsense capabilities by boosting the performance on WS via simple probability ranking (Trinh and Le, 2018; Brown et al., 2020; Zhou et al., 2020). Another advancement was the curation of a large, crowdsourced dataset for WS, Winogrande (Sakaguchi et al., 2019). Models that train on this dataset are close to human performance. But are we any closer to achieving commonsense reasoning?

We provide three explanations for the perceived progress on the WS task: (1) lax evaluation criteria; (2) biases in the datasets that remain despite efforts to remove them, and (3) knowledge and reasoning leakage from large training data. After proposing stricter evaluation criteria (§3) and exposing the existence of superficial hints and quantifying their magnitude using control baselines (§5), we pro-

³There has been some theoretical work that analyzed WS sentences and proposed a framework, the “correlation calculus,” arguing that resolving these problems involves a discourse coherence (Bailey et al., 2015; Michael, 2015).

pose to evaluate the commonsense knowledge in the PLMs directly, without exposing them to the supervised training set (§6); we show that on all the models we tested, the performance on the task is random. Examples for WS, the proposed control baselines, and zero-shot instances can be found in Figure 1.

Our main premise in this work is that, from a commonsense perspective, the generalization capabilities models can get from large training data are limited. Due to the vast number of commonsense facts (e.g. steel is a hard material, planets are huge, etc.), it is infeasible to learn them all from a limited-scale training set. However, this knowledge can still be acquired in different ways, such as self-supervision (Mitchell et al., 2015), Open IE (Tandon et al., 2014), collecting statistics from large text corpora (Elazar et al., 2019), PLMs Zhou et al. (2020) and more (Bagherinezhad et al., 2016; Forbes and Choi, 2017). Therefore, we claim that the vast majority of commonsense knowledge a model has should come from sources external to the supervised dataset. The supervised training set should mainly provide a means for learning the format of the task but not as a source for commonsense knowledge acquisition. We thus question the approach, which has recently gained popularity (Sakaguchi et al., 2019; Klein and Nabi, 2020), of using models trained on large datasets for evaluating general commonsense reasoning capabilities, like WS.

Contributions. (i) We begin by proposing a general evaluation method that makes use of groups that contain similar inputs, e.g. the twin sentences in WS (§3). That is, instead of measuring accuracy by scoring each sentence separately, we suggest scoring according to the worse score on both inputs: giving a point only if both sentences are predicted correctly. This evaluation reduces the risk of successful prediction due to biased examples and better reflects the models’ commonsense reasoning abilities. (ii) Next, we extend previous work (Trichelair et al., 2019) that found in the Winograd Schema Challenge (WSC)—through manual evaluation—biased examples that can be solved using simple statistics. We propose two automatically constructed control baselines that distort the sentences to be nonsensical, on which a score higher than majority suggests the presence of biased examples (§5). We find that WSC (Levesque et al., 2012) is highly biased, whereas the newly

suggested dataset, WinoGrande (Sakaguchi et al., 2019), is less biased.⁴

(iii) Finally, to bypass the supervised training step, we propose to directly evaluate PLMs on WS in a zero-shot setup; this allows for assessing how many commonsense reasoning capabilities were acquired in the pretraining step. Specifically, this evaluation disentangles the commonsense capabilities of PLMs from the knowledge they acquire from the training set. Combining our new evaluation method and taking into account the biases in the data with the zero-shot setting, we show that all models we test perform randomly in the zero-shot setting. We then demonstrate using learning curves of models trained on increasing amounts of data, that it takes huge amounts of training instances to make small improvements in the test set, demonstrating the ineffectiveness of large training sets in acquiring commonsense reasoning skills. We interpret these results as evidence that a lot of the commonsense reasoning capabilities are learned during fine-tuning, as opposed to the pre-training step.

Based on our experiments, we conclude that many of the claims of progress on WS in recent years are unjustified, and stem from sub-optimal evaluation, biases, and commonsense knowledge learned from a supervised training set. Nevertheless, we suggest that the newly proposed Wino-grande dataset (Sakaguchi et al., 2019) shouldn’t be used for training, but it provides good data for evaluation, and hope that our new evaluation methods will assist faithful tracking of commonsense reasoning progress.

2 Background

2.1 WSC and the Twin Sentences

The Winograd Schema (Levesque et al., 2012) was constructed to serve as a benchmark for commonsense reasoning capabilities of models (similarly to the way Textual Entailment was proposed to serve as a benchmark for entailment capabilities of models (Dagan et al., 2005, 2013)). WSC contains a small test set of 273 examples, created by experts, and for several years models were struggling to perform well on it. Each question involves four key features: 1) two entities are mentioned in each sentence, and they can be two males, two females,

⁴In Appendix D, we provide details on how AFLITE, the algorithm that was used to filter examples from Winogrande operates, and how it is different from our baselines.

two inanimate objects, or two groups of people or objects; 2) a pronoun or a possessive adjective is used in the example to refer to one of the entities; 3) the task is to determine which of the two entities is referred to by the pronoun, and 4) each sentence contains a *special word* which, when replaced, the answer changes. There are no other limitations on the sentences besides these constraints and, consequently, this test is considered to be a general commonsense reasoning test, unlike other benchmarks, which focus on specific commonsense capabilities (Rashkin et al., 2018; Forbes et al., 2019; Sap et al., 2019a,b; Bisk et al., 2020).

In order to fulfil the fourth feature, each example was paired with an additional *twin* sentence, which only slightly differs from its twin. (Similar tests sets were recently proposed and referred to as *Counterfactual data* (Kaushik et al., 2019) and *Contrast sets* (Gardner et al., 2020)). For example, the *twin* sentence of Example 1 is:

2. *The trophy* would not fit in *the brown suitcase* because **it** was too small.

Notice that the special words in these sentences are *large* and *small*, and in this sentence, **it** refers to *the brown suitcase* (as opposed to *the trophy* in Example 1). The special word is a key part of WS, that makes the task hard to solve. These words were chosen carefully to avoid statistical correlations between the special word and the entities. In this example, both *trophy* and *suitcase* can be small, which makes the task hard to solve by machines; and as Levesque et al. puts it: “This helps make the test Google-proof: having access to a large corpus of English text would likely not help much (assuming, that answers to the questions have not yet been posted on the Web, that is)!”

2.2 Progress on WSC

Since WSC was proposed as a benchmark for commonsense (Levesque et al., 2012), there were many attempts to improve performance on this benchmark, that involved different approaches from web queries (Rahman and Ng, 2012; Sharma et al., 2015; Emami et al., 2018), using external knowledge sources (Sharma, 2019), information extraction and reasoning (Isaak and Michael, 2016) and more (Peng et al., 2015; Liu et al., 2017a,b; Fähndrich et al., 2018; Klein and Nabi, 2019; Zhang et al., 2019, 2020a).

Newer approaches use LMs to assign a probability to a sentence by replacing the pronoun with

an entity, one at a time, and pick the more probable sentence (Trinh and Le, 2018; Opitz and Frank, 2018; Radford et al., 2019; Kocijan et al., 2019). More recently, sequence to sequence models have been employed to directly predict the referred entity using training (Raffel et al., 2020), or in a zero-shot or few-shot setting citebrown2020language. The latest results of GPT-3 (Brown et al., 2020) are rather impressive, and agree with the premise of this paper, as the model sees none to a few dozen examples to learn the format. It is worth noting, though, that the training corpus of GPT-3 included some of the WSC questions, and therefore these results should be taken with a grain of salt. For a comprehensive review of the progress on approaches and related datasets of WS, see Kocijan et al. (2020).

Zhou et al. (2020) probed multiple LMs for commonsense capabilities in different datasets among WSC, by computing the probability the LM assigns each alternative and choosing the more probable one. The advantage of this method is its unsupervised approach; it does not teach the model any new knowledge. Notably, their evaluation protocol, which computes the average log probability of each masked word is problematic, since special words that get tokenized into more than one word-piece are still masked independently, thus priming the model towards a certain answer. In this work, we propose a new evaluation methodology and show that these models’ performance is random. Finally, Zhang et al. (2020b) provided an analysis of different types of commonsense knowledge needed to solve the different WSC questions, including properties, eventualities, and quantities.

3 A Robust Group Score Evaluation

Many works in recent years have shown that large neural networks can achieve high performance on different benchmarks while “being right for the wrong reasons” (McCoy et al., 2019). These successes arise from a variety of reasons such as biases in datasets (Poliak et al., 2018; Tsuchiya, 2018; Gururangan et al., 2018; Kaushik and Lipton, 2018), annotators biases (Geva et al., 2019), etc. Levesque et al. (2012) proposed to alleviate some of these issues by using the twin sentences along with the special word. However, the proposed evaluation of WSC still scored each twin separately. As Trichelair et al. (2019) showed that some WSC instances can be solved using simple correlations,

we argue that the independent scoring may result in unjustifiably inflated scores. Here, we inspect a new evaluation that accounts for some of these biases and provide a more robust evaluation for cases where we have grouped instances (e.g. minimal pairs).

3.1 Group Scoring

Recent studies proposed to augment test instances with minimal pairs, that either change the original answer (Kaushik et al., 2019; Gardner et al., 2020), or keep it intact by using paraphrasing, synonyms, etc. (Glockner et al., 2018; Shah et al., 2019). Typically, these works report the results separately on a new test set.

We extend over previous work that proposes to evaluate pairs (Abdou et al., 2020) or groups (Elazar et al., 2021) of related instances and assign a point only if all are correctly predicted by a model. Our evaluation framework exploits groups of minimal-distance instances and results in a more robust evaluation. Specifically, for an arbitrary scoring function f , and a group of minimal-distance instances x_i , score each of the examples x_{i_j} in the group and assign the group its worse-performing score.⁵

$$groupScore(x_i) = \min_j f(x_{i_j})$$

The motivation behind this new evaluation is three-fold: (1) Predicting correctly all examples in a group provides a more robust measurement, and indicates a better understanding of the instances; (2) The lowest scored example is the groups’ “Achilles heel” and thus makes the success on other examples suspicious; (3) It lowers the probability of random predictions (especially in classification tasks), or the use of shallow heuristics to solve examples. We note that cases where all examples in a group can be solved based on some bias will still lead to a high score on this group. Therefore this evaluation does not solve the problem of biased examples, but it reduces the chance of scoring them as correct in cases where not all the groups’ instances are biased.⁶

In classification tasks, a consequence of this evaluation is the change in random performance. For example, in the case of balanced binary classification, the chance accuracy drops from 50% to 25%.

⁵The minimum in cases where higher scores indicate better performance, and maximum otherwise.

⁶A similar evaluation was used by Zhou et al. (2019), with the “Exact Match” metric for a multi-label classification task.

This generic evaluation can be applied not only in classification tasks but also in other tasks that use different evaluation metrics such as BLEU and ROUGE in generation (Papineni et al., 2002; Lin, 2004). For WS, we use *group scoring* over the twin sentences, with accuracy as the per-instance scoring function. This yields the paired evaluation that was recently proposed by Abdou et al. (2020) for evaluating WSC.

3.2 Other Robust Evaluation Protocols

It is important to note that any WS test set is only an approximation of the commonsense reasoning skills required overall. The twin-sentences allow to test for specific skills (such as the interchange between small and large with ‘fit’ in Examples 1, 2), but other perturbations are possible which allow testing different skills. For instance, Abdou et al. (2020) proposed several perturbations on the original sentences that mostly do not change the answer, such as synonyms entities substitution, tense switch, gender switch, etc. These perturbations are also reminiscent of the *switched* protocol of Trichelair et al. (2019), where models are evaluated on examples where the candidates can be switched in the order (which mainly happens with proper names, but also with inanimate objects), expecting a consistent prediction from models since the label does not depend on the entities order. Under the *group-scoring* evaluation, we expect a model to succeed on all perturbations from the same group.

4 Setup

Datasets We experiment with two English WS datasets:

Winograd Schema Challenge (WSC) (Levesque et al., 2012) contains 273 manually curated examples. We also report results on the debiased (or, *non-associative*) examples that were filtered by Trichelair et al. (2019), named WSC-na.

Winogrande (Sakaguchi et al., 2019) is a recent crowdsourced dataset that contains WS questions. Winogrande contains 40,938, 1,267, 1,767 examples for train, development, and test respectively. Since the test labels were not published, we report our results on the development set.

Modeling We follow the modeling of Sakaguchi et al. (2019), which finetunes PLMs as a multiple-choice problem on the Winogrande training set. In this modeling, the pronoun is replaced with either

Dataset	Setup	Single	Group
WSC	original	89.71	79.41
	<i>no-cands</i>	60.72	40.35
	<i>part-sent</i>	64.88	33.88
WSC-na	original	89.45	79.09
	<i>no-cands</i>	58.06	34.41
	<i>part-sent</i>	59.90	25.00
Winogrande	original	71.49	58.45
	<i>no-cands</i>	53.07	31.05
	<i>part-sent</i>	53.11	22.34

Table 1: Results of RoBERTa-large trained on Winogrande, evaluated on the different datasets in the regular condition (original) and the two bias-exposing baselines. Reporting results both on the original accuracy (Single), and the group-scoring (Group). Random performance on the single and group-scoring evaluations are 50% and 25% respectively.

one of the entities, and the ‘[CLS]’ token representation is used for prediction. We also experiment with another loss for WS, that was explored in Liu et al. (2020) where instead of using a different classification head, uses the original MLM head for predictions. We report these results in Appendix G.

Pre-trained Models We experiment with multiple PLMs: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019).

5 Biases-Detecting Baselines for WS

WSC was carefully designed by human experts to minimize the creation of biased examples. For instance, Example 3 is not considered as a good WS example since a *racecar* is more likely to *go fast* rather than a *school bus*.

3. The *racecar* zoomed by the *school bus* because **it** was going so fast.

However, such correlations are often easy to miss. As evidence, Trichelair et al. (2019) found 37 sentences to be *associative*, or *non Google-proof*.⁷ These examples were labeled manually using crowdsourcing, therefore these are still bound to what non-experts can catch, and subtler cues may be hard to catch. Other correlations may be harder or impossible to be detected by humans since they

⁷*Google-proof* is an attribute introduced in Levesque et al. (2012) that refers to the strength of a test, and the inability to solve it by having access to a large text corpora.

are the result of spurious correlations. These features, which can be learned during pretraining or fine-tuning, may result in successful predictions that do not reflect commonsense reasoning skills.

To account for these biases, we propose two *control baselines*, which are likely to achieve random performance with an unbiased model. A score above random indicates the presence of biased examples.

No-Candidate Baseline This baseline (*no-cands*) removes the two candidates (entities) from the text. For instance, Example 1 will turn into: “would not fit in because **it** was too large.”

Partial-Sentence Baseline In this baseline (*part-sent*) we split the sentence into two parts, based on punctuation and discourse markers⁸ and take only the part containing the pronoun. For instance, Example 1 will be transformed into the following: “because **it** was too large.” A similar approach was used by Trichelair et al. (2019), however, they employed annotators to manually indicate whether the partial sentence containing the pronoun is associative to one of the candidates. Alternatively, we use a trained model and inspect the overall score on a controlled dataset. We note that these two baselines create nonsensical sentences. Therefore, we expect humans to not be able to properly solve them. Thus, a model that achieves higher than random performance on these baselines over a large enough dataset is suspected to rely on spurious correlations.

These baselines are reminiscent of previous works that used part of the input (e.g. the hypothesis only baseline in NLI), to reveal biases in multiple NLI datasets (Poliak et al., 2018) and reading comprehension (Kaushik and Lipton, 2018).

5.1 Results

We retrain the RoBERTa large model from Sakaguchi et al. (2019) that was trained on Winogrande and report the results using the original and the new group-based evaluations in Table 1. On WSC this model achieves 89.71% and 79.41% accuracy, on WSC-na it achieves 89.45% and 79.09%, and on the dev set of Winogrande, it achieves 71.49% and 58.45% accuracy, respectively. To make these evaluations comparable, we filter sentences with no twin sentences from Winogrande and the single

⁸‘so’, ‘but’, ‘and’, ‘because’, ‘although’, ‘though’, ‘due’, ‘since’, ‘;’, ‘,’, ‘;’, ‘?’

ID	WSC	Trichelair et al.	no-cands	part-sent
2	The <i>trophy</i> doesn't fit into the brown <i>suitcase</i> because it is too large.	✗	✓	✓
8	The <i>lawyer</i> asked the <i>witness</i> a question, but he was reluctant to <u>repeat</u> it.	✗	✗	✗
53	The <i>painting</i> in Mark's living room shows an <i>oak tree</i> . it is to the right of a <u>house</u> .	✓	✗	✗
54	There is a <i>gap</i> in the <i>wall</i> . You can see the garden <u>through</u> it .	✓	✓	✓

Table 2: Instances from WSC, along with indication if the manual filtering by Trichelair et al. (2019) marked them as associative, as well as our proposed baselines correctly predicted by group-scoring.

triplet sentence from WSC, remaining with 568 and 272 instances, respectively (or, 284 and 136 pairs). The resulting performance on the original Winogrande development set is 78.3%.⁹ The single accuracy score on sentences that have pairs is lower by almost 8 points than the original set, which suggests that the sentences with no pair are easier, and may contain some biases. Next, we highlight the performance difference between the original evaluation and the paired, which dropped by 10.3, 10.36, and 13 points for WSC, WSC-na, and Winogrande, respectively. Finally, the results on our proposed baselines achieve higher performance than the random baseline, especially for WSC. The *no-cands* baseline achieves 40.35%, 34.41%, and 31.05% on WSC, WSC-na, and Winogrande respectively, whereas the *part-sent* baseline achieves 33.88%, 25.00%, and 22.34% accuracy. These results indicate that WSC is highly biased (over 15 points above random performance), and even after the manual filtering of Trichelair et al. (2019) some statistical correlations remain. On Winogrande, the no-cands baseline achieves more than 6 points above random, but is less biased than WSC and WSC-na, presumably due to the AFLITE algorithm.

5.2 Qualitative Analysis

In Table 2 we inspect some instances from WSC and indicate if the manual filtering by Trichelair et al. (2019) found them to be biased, as well as successful predictions on the twin sentences by our baselines. Although successful predictions may result from chance (although the probability that both baselines correctly predicted both pairs is relatively low - 6.25%), we highlight some cases we find interesting.

The first example from the table was predicted correctly by both our baselines, but not by Trichelair et al. (2019). This may be a case of memorization of this very popular example, by the

pretrained RoBERTa model which was trained on many webpages (Emami et al., 2020). We provide more compelling evidence for this example's memorization in Appendix F. Example ID 54 was predicted correctly by our baselines, as well as by Trichelair et al. (2019) since this example is associative: it is likely to see something through a gap, rather than through a wall.

6 Disentanglement of Commonsense Reasoning and Learned Commonsense

In this section, we wish to disentangle the commonsense reasoning skills acquired by PLMs, and what they learn during fine-tuning on a WS dataset. We propose a method that allows evaluating pretrained Masked Language Models (MLM) in a zero-shot setting on WS-like questions.

6.1 Zero Shot MLM Evaluation

Previous work proposed to evaluate MLMs in a zero-shot setting by replacing the pronoun with masked tokens, corresponding to the number of tokens the entities are tokenized into. Then, by inspecting each entity's probability the more probable entity is selected (Kocijan et al., 2019; Abdou et al., 2020). However, this approach is problematic when the entities are of different token lengths or consist of more than a single token since the model may be primed towards a certain answer. For instance, consider Example 1's entities, trophy and suitcase, in the case they are tokenized into *trophy* and *suit, case*. In this scenario, the MLM will see in one of the cases a single mask (assigning the *trophy* probability), and in the other case, it will see two masks (assigning the *suit* and *case* probabilities). Since the model has access to the number of tokens it has to complete, the comparison between these two options is flawed. Another approach, used by Zhou et al. (2020) is to calculate a probability to the entire sentence, by masking a single token at a time. However, this method is also problematic when the entities are tokenized into more than a sin-

⁹Compared to 79.3%, reported by Sakaguchi et al. (2019).

Model	WSC		WSC-na		WinoGrande	
	Single	Group	Single	Group	Single	Group
random	50.00	25.00	50.00	25.00	50.00	25.00
BERT-base	56.52	15.22	54.79	12.33	53.12	11.11
BERT-large	61.41	23.91	60.27	21.92	55.56	12.50
RoBERTa-base	63.04	27.17	60.27	21.92	56.25	14.58
RoBERTa-large	73.91	47.83	71.23	42.47	54.86	12.50
ALBERT-base	55.43	13.04	55.48	12.33	52.78	7.64
ALBERT-xxlarge	78.80	57.61	77.40	54.79	58.68	20.83

Table 3: Performance of different PLMs evaluated in the zero-shot setup of WS. Single refers to the standard accuracy over the entire test set, Group refers to group-scoring.

gle token since unmasked-tokens are affecting the prediction of the masked tokens. For instance, following the same example as before, where *suitcase* is tokenized into *suit* and *case*, a model that sees *suit* is more likely to assign a high probability to *case*, therefore staining the probability distribution, and causing a wrong comparison.

Since properly evaluating MLM on WS sentences with more than a single word that differs between the sentences is challenging, we filter these examples. Then, we mask this word, and compare the probabilities of the two candidates, as was done in previous work (Goldberg, 2019; Talmor et al., 2019; Ettinger, 2020). The issue with this approach is that typically, the candidates are constructed of multiple words or named entities that are often tokenized into multiple word-pieces, which will result in filtering a great portion of the data. Instead, we propose to make use of the *special* word (the word that is different between the twin sentences), mask it, and replace the pronoun with the correct answer. Then, the model has to decide which of the special words refers to each entity. Occasionally, there is more than one special word, or it gets tokenized into multiple tokens, therefore we discard these sentences.

An example of this transformation process on Example 1 is the following:

4. *The trophy* would not fit in *the brown suitcase* because the trophy was too [MASK].

where ‘[MASK]’ is the token that has to be predicted between the two original special words: ‘large’ or ‘small’. The twin sentence of this example would accordingly be the same but with the entity ‘the trophy’ replaced with ‘the brown suitcase’, and the correct answer would change from ‘large’ to ‘small’.

One potential pitfall of this formulation is that it is not faithful to the original WS, and tests a different mechanism. To test the difference between these formulations, we train the RoBERTa large model on Winogrande on our transformed data, and compare it to the results of the same model, trained on the original setup. We make sure to only use sentences that can be transformed, assuring to train both models on the same subset. The model’s performance on the original setup achieves 66.10% and 55.93% on the original and paired evaluation development set, whereas the model trained on the transformed setup achieves 70.06% and 64.97%. The latter achieves higher performance, suggesting that our transformation may be preferable in modeling, or easier than the original setup. Since this modeling is easier for the model, the results provide a higher bound of the original results, making the results even more alarming.

We transform WSC, WSC-na, and the Wino-grande dev set with the proposed method and remain with 226, 180, and 354 examples, respectively. We then evaluate the pre-trained LMs described in Section 4, and report the results in Table 3. We note that the overall performance is much lower compared to the finetuned model, as expected. Next, the performance on the group-scoring on WSC-na is relatively low, except for RoBERTa-large and ALBERT-xxlarge, which achieve 42.47 and 54.79, high above random performance. On the other hand, the performance on Winogrande, across all models is below random performance (best result by ALBERT-xxlarge, of 20.83%), indicating poor commonsense capabilities of these models. Since we found in the previous Section (§5) that WSC and WSC-na are highly biased, we take the results on Winogrande to better reflect commonsense reasoning skills. Recall

that the comparison between the two formulations suggested that our new formulation should perform better, a fact that makes the random predictions in the zero-shot setup even more remarkable.

7 Progress in Commonsense Reasoning?

The large performance gap may not seem surprising. In most tasks in NLP, we do not expect a PLM to do well on new tasks out of the box and expect a supervised dataset to provide the required skills. However, we claim that for commonsense tasks, this argument does not hold. Since commonsense reasoning skills and knowledge are huge, it is not likely to acquire all that information through supervision. Consider the following WS instances:

5. *The large ball crashed right through the table because it was made of steel.*
6. *I bought a steel property at the same time as my wooden property. The _ property was harder.*

Examples 5 and 6¹⁰ come from WSC and Winogrande training set, respectively. The fact that steel is a strong material is part of the knowledge needed to solve Example 5. However, a model that is trained on Example 6 may pick up this fact. Will this training instance also teach the model facts about other materials, such as *styrofoam*?

To quantify the effect of training on the success in solving WS questions, we re-split Winogrande training set to leave enough data for testing (out of the ‘unbiased’ subset) and use the rest for training. From the remaining training set, we create multiple training splits, increasing in size, to study the effect of increasing amounts of data on the overall performance. We use the original development set to pick the best models. We report learning curves with the different models, where each point is the average score of three runs, in Figure 2.¹¹ We report the number of correct pairs predicted correctly on the y-axis as a function of the training size. These curves indicate that the inspected models obtained no commonsense reasoning capabilities in the pretraining step, and are slowly improving their performance the more data they are trained on. However, except for a sudden improvement with 500 examples for ALBERT, the

¹⁰Winogrande was collected with ‘_’ instead of pronouns.

¹¹Full numeric results, along with standard deviations are reported in Appendix C.

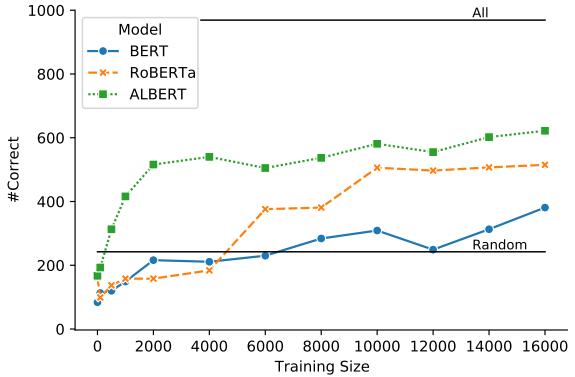


Figure 2: Learning curves for the large versions of BERT, RoBERTa, and ALBERT models, trained on increasing amounts of data. This figure differs from the Winogrande leaderboard. We explain the source of these differences in Appendix E.

slope increases incredibly slowly and requires a significant amount of additional training instances for small improvements (BERT and RoBERTa’s slopes are more moderate). We conclude that training data is mostly non-beneficial for commonsense reasoning, and models should acquire it using other methods.

We note that the initial fast increase in ALBERT’s performance is interesting, and may be due to another explanation; that is commonsense reasoning is composed of commonsense knowledge (e.g. steel is hard), and reasoning (comparing between objects sizes). Some of the knowledge may be encoded in these models, and reasoning can be taught. However, if that’s the case, datasets should account for that, with careful splits. We leave the answer to this question to future work. Overall, this increase is nevertheless rather moderate, and once a model passes this point (about 500 examples), the performance increases slowly, which goes in line with our claims.

8 Conclusions

In this work, we begin by discussing the current evaluation of WS and propose an additional evaluation metric, *group-scoring*, that credits a model with the worse performing instance of a group. While in this work we focus solely on WS, we propose to use the same evaluation also in other domains, where minimal pairs are available (Kaushik et al., 2019; Gardner et al., 2020; Warstadt et al., 2020), as a more reliable evaluation metric. We then propose two new control baselines that account for biases in WS data and show that WSC is

highly biased, while Winogrande shows much less biased data.

Finally, we propose a method to evaluate MLMs on WS sentences in a zero-shot setting. We show that the performance of popular MLMs is random and that models improve gradually the more training data they see. We conclude that the use of large training sets is not always desirable, especially in commonsense reasoning settings, and call future work to find other methods to improve our models' commonsense abilities.

Acknowledgements

We would like to thank Vered Shwartz, Keisuke Sakaguchi, Rotem Dror, Niket Tandon for helpful discussions and comments on early versions of this paper.

References

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. *The sensitivity of language models and humans to Winograd schema perturbations*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7590–7604, Online. Association for Computational Linguistics.

Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2016. Are elephants bigger than butterflies? reasoning about sizes of objects. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3449–3456.

Daniel Bailey, Amelia J Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. 2015. The winograd schema challenge and reasoning about correlation. In *AAAI Spring Symposia*. Citeseer.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, pages 7432–7439.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Morgan and Claypool.

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.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, E. Hovy, Hinrich Schütze, and Y. Goldberg. 2021. Measuring and improving consistency in pretrained language models. *ArXiv*, abs/2102.01017.

Yanai Elazar, Abhijit Mahabal, Deepak Ramachandran, Tania Bedrax-Weiss, and Dan Roth. 2019. How large are lions? inducing distributions over quantitative attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3973–3983, Florence, Italy. Association for Computational Linguistics.

Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. A knowledge hunting framework for common sense reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1949–1958.

Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2020. An analysis of dataset overlap on winograd-style tasks. *arXiv preprint arXiv:2011.04767*.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Johannes Fähndrich, Sabine Weber, and Hannes Kanthak. 2018. A marker passing approach to winograd schemas. In *Joint International Semantic Technology Conference*, pages 165–181. Springer.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 266–276.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

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

via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

M. Geva, Y. Goldberg, and J. Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. **Breaking NLI systems with sentences that require simple lexical inferences.** In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

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

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.

Nicos Isaak and Loizos Michael. 2016. Tackling the winograd schema challenge through machine logical inferences. In *STAIRS*, volume 284, pages 75–86.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tassilo Klein and Moin Nabi. 2019. **Attention is (not) all you need for commonsense reasoning.** In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4831–4836, Florence, Italy. Association for Computational Linguistics.

Tassilo Klein and Moin Nabi. 2020. **Contrastive self-supervised learning for commonsense reasoning.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7517–7523, Online. Association for Computational Linguistics.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. **A surprisingly robust trick for the Winograd schema challenge.** In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842, Florence, Italy. Association for Computational Linguistics.

Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. 2020. A review of winograd schema challenge datasets and approaches. *arXiv preprint arXiv:2004.13831*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haokun Liu, William Huang, Dhara Mungra, and Samuel Bowman. 2020. Precise task formalization matters in winograd schema evaluations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8275–8280.

Quan Liu, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2017a. Cause-effect knowledge acquisition and neural association model for solving a set of winograd schema problems. In *IJCAI*, pages 2344–2350.

Quan Liu, Hui Jiang, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2017b. Combing context and commonsense knowledge through neural networks for solving winograd schema problems. In *AAAI Spring Symposia*.

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*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. 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.

Julian Michael. 2015. The theory of correlation formulas and their application to discourse coherence. Bachelor’s Thesis.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

Juri Opitz and Anette Frank. 2018. Addressing the winograd schema challenge as a sequence ranking task. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 41–52.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473.

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

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. 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 4453–4463.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6649–6658.

Arpit Sharma. 2019. Using answer set programming for commonsense reasoning in the winograd schema challenge. *Theory and Practice of Logic Programming*, 19(5-6):1021–1037.

Arpit Sharma, Nguyen Ha Vo, Somak Aditya, and Chitta Baral. 2015. Towards addressing the winograd schema challenge-building and using a semantic parser and a knowledge hunting module. In *IJCAI*, pages 1319–1325.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olympics-on what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

Niket Tandon, Gerard De Melo, and Gerhard Weikum. 2014. Acquiring comparative commonsense knowledge from the web. In *AAAI*, pages 166–172.

Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the winograd schema challenge and swag. 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 3373–3378.

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. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mo-hananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pier-ric 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.

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019. Sp-10k: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 722–731.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020a. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20–24, 2020*, pages 201–211.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020b. **WinoWhy: A deep diagnosis of essential commonsense knowledge for answering Winograd schema challenge**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745, Online. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. 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 3354–3360.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *AAAI*, pages 9733–9740.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*.

A Detailed Setup

Datasets We report our results on two datasets:

Winograd Schema Challenge (WSC) (Levesque et al., 2012) contains 273 manually curated examples. Each example is paired with a *twin-sentence*, meaning that there’s a special word that is changed between the two sentences, that changes the coreferring entity. Trichelair et al. (2019) have labeled the original WSC examples, and found 37 examples to be biased (or, *associative* Trichelair et al. (2019)). We thus also use the ‘unbiased’ subset, named *non-associative* which excludes the associative examples. We refer to this subset as WSC-na

Winogrande (Sakaguchi et al., 2019) is a recent crowdsourced dataset that contains WS questions. Winogrande is much larger than WSC and contains 9,248, 1,267, 1,767 examples for train, development, and test respectively. Winogrande was filtered from ‘biases’ using their proposed AFLITE algorithm, which produced the mentioned challenging dataset. However, the authors also release and use the ‘biased’ instances for training, making a total of 40,938 training instances.

Pre-trained Models We experiments with multiple pre-trained models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019). These models are large Transformer-based architectures (Vaswani et al., 2017), that are trained on the Masked Language Modeling task, which is predicting the masked word in a given context. These models are pretrained on huge amounts of text such as Wikipedia, the book corpus (Zhu et al., 2015), parts of CommonCrawl, and more. Specifically, we conduct our experiments with BERT-large-cased, RoBERT-Large, and ALBERT-XXLarge-V2, which have 335M, 335M, and 223M parameters, respectively.

B Implementation Details

We implemented the experiments with the huggingface package (Wolf et al., 2020). Following the previous work (Sakaguchi et al., 2019), on all our experiments, we set the learning rate to be 1e-5, batch size to be 8, and trained the models for 8 epochs. Adam (Kingma and Ba, 2015) is used as the optimizer. We trained our model with RTX 2080, and the training time is 13, 14, and 62 minutes per epoch on the largest training set Winogrande (10) for BERT-large, RoBERTa-large, and Albert-XXL-v2, respectively. As the evaluation is

conducted on the dev set, we do not use it to select the best model. Instead, we report the performance with the final model, which is converged based on our observation.

C Full Learning Curves Results

The full results from Figure 2, along with the standard deviations, are reported in Table 4.

D AFLITE Details

AFLITE (Sakaguchi et al., 2019), an algorithm proposed for reducing datasets’ biases was used to create Winogrande (Sakaguchi et al., 2019). It works as follows: a RoBERTa model (Liu et al., 2019) is finetuned on a random subset of the data to train a ‘weak’ model of the task. Then, the rest of the instances are encoded using the model’s encoder. Then, for multiple iterations, a set of weak classifiers (linear) are trained on a subset of the encoded data and predict the rest. If more than k classifier predicted correctly an instance’s label, it is discarded from the final dataset. This process is repeated multiple times, until reaching a satisfying dataset size (which is controlled by predefined hyperparameters).

Although this algorithm filter examples that are ‘easy’, as a set of linear models that were trained on a medium quality representation managed to predict the correct answer, it is unclear how unbiased the dataset is. In contrast, our proposed baseline methods directly detect biases the classification model may rely on, by presenting challenging perturbations on which a model is not likely to succeed above random. Thus, our procedure is inherently different than the general-purpose AFLITE filtering algorithm.

E Comparison to Winogrande Leaderboard

We note that Figure 2 differs from the Winogrande leaderboard in multiple ways: first, we compare different models than the ones that appear on the leaderboard. Specifically, the to-date leading submission (accurate as of March 21st, 2021), UNICORN, does not provide details about the model, except it is a T5 based model, trained on a collection of datasets. Since the content of these datasets is not publicly available, it is impossible to assess the quality of this submission. For instance, if one of these datasets contains other commonsense reasoning datasets, the model may have picked up

# Training	BERT		RoBERTa		ALBERT	
	Single	Group	Single	Group	Single	Group
0	52.99 (0.00)	8.67 (0.00)	56.39 (0.00)	16.61 (0.00)	55.55 (0.00)	17.23 (0.00)
100	53.47 (0.75)	11.71 (0.75)	52.78 (0.75)	10.22 (4.48)	58.24 (1.49)	19.89 (2.74)
500	49.31 (1.87)	12.42 (1.74)	49.65 (0.50)	14.17 (1.99)	60.07 (0.50)	32.35 (1.27)
1,000	51.39 (0.99)	15.33 (0.75)	50.35 (0.37)	16.33 (0.50)	62.50 (0.62)	42.89 (0.25)
2,000	51.39 (0.87)	22.32 (3.49)	49.65 (0.25)	16.35 (1.49)	62.85 (2.36)	53.27 (3.24)
4,000	48.96 (2.61)	21.73 (2.49)	49.65 (0.50)	18.94 (1.24)	67.36 (1.12)	55.72 (2.49)
6,000	50.35 (0.50)	23.73 (0.50)	59.72 (1.86)	38.85 (2.99)	67.71 (2.86)	52.16 (3.73)
8,000	48.26 (1.24)	29.27 (0.75)	50.35 (1.37)	39.32 (1.99)	67.36 (0.50)	55.43 (0.25)
10,000	51.39 (1.12)	31.85 (1.99)	62.85 (0.12)	52.27 (0.99)	73.76 (1.76)	59.98 (1.94)
12,000	50.00 (1.62)	25.68 (1.49)	62.85 (0.50)	51.24 (0.50)	72.22 (1.33)	57.28 (0.54)
14,000	52.08 (0.50)	32.31 (3.24)	62.15 (1.49)	52.31 (0.75)	75.61 (0.63)	62.15 (2.24)
16,000	54.86 (0.75)	39.31 (1.99)	60.42 (2.11)	53.14 (3.24)	76.82 (1.15)	64.21 (1.42)

Table 4: Effect of the training data size on different models performance. We report results on BERT, RoBERTa and ALBERT, all with their largest variants.

on commonsense reasoning skills which are also tested for in Winogrande. Second, the leaderboard uses the original evaluation, based on the accuracy of single instances. As we claim in Section 3, this evaluation is sub-optimal and causes an over-estimation of the actual performance of models. Moreover, our analyses were done on the development set, as opposed to the reported test set performance, since the test set is not publicly available. Finally, the leaderboard presents a learning curve of 5 training sizes, as we report the results over 12 different training sizes.

F Elaborate Analysis

In Section 5.2 we showcase some examples from WSC and provide possible explanations for which our baselines (§5) are able to solve them. Here, we provide additional evidence that supports our claim. We do so for the example where both baselines predict the correct answer, but the manual inspection from Trichelair et al. (2019) does not consider it to be biased. We emphasize that this example is not biased per se, and thus the annotation from Trichelair et al. (2019) was correct, but the pretrained model, which was trained on the web, may have caught up statistical cues that help it predict these examples correctly, even with partial information. For completeness, we repeat the example here:

7. The *trophy* doesn’t fit into the brown *suitcase* because **it** is too large.

Example 7 is a popular example that is often given when describing the task in the media. As

evidence, we search for this sentence in Google and found it in multiple websites:

- <https://theness.com/neurologicablog/index.php/a-tougher-turing-test/>
- <https://www.eitdigital.eu/newsroom/blog/article/whats-too-big-the-trophy-or-the-suitcase/>
- <https://cmte.ieee.org/futuredirections/2014/08/20/whats-too-big-the-trophy-or-the-suitcase/>

Next, we search for these websites in Common Crawl¹², the February 2019 version that was reported to be part of RoBERTa’s training data (Liu et al., 2019). We use an index server¹³ that allows querying a specific index and look specific websites. We find that the first two websites are included in this index. Although we cannot guarantee that these websites were part of RoBERTa’s training data since it was not published, the probability that several examples from WSC were part of the large training data of RoBERTa (and later models), with these websites, or other, is high.

Dataset	Setup	Single	Group
WSC	original	89.71	80.88
	<i>no-cands</i>	60.96	29.82
	<i>part-sent</i>	59.09	22.31
WSC-na	original	90.00	81.82
	<i>no-cands</i>	59.14	25.81
	<i>part-sent</i>	56.77	16.67
Winogrande	original	70.95	54.23
	<i>no-cands</i>	54.87	17.69
	<i>part-sent</i>	54.43	14.18

Table 5: Results of RoBERTa-large trained on Winogrande, evaluated on the different datasets in the regular condition (original) and the two bias-exposing baselines using the MC-MLM loss (Liu et al., 2020). Reporting results both on the original accuracy (Single), and the group-scoring (Group). Random performance on the single and group-scoring evaluations are 50% and 25% respectively.

G MLM results

Here we report the results for the MC-MLM loss that was explored in Liu et al. (2020), where instead of training a new head for the classification task, it uses the original MLM head and scores the different candidates instead of the pronoun. We run all experiments including fine-tuning, and report the results in this section.

First, the bias experiment results are detailed in Table 5. Although the results on the standard setting (*original*) are similar to the ones when using a dedicated head (Table 1), this model appears to be less biased: the *no-cands* baseline still perform better than random on WSC, but the other baseline and the other evaluations perform randomly.

Finally, we repeat the learning curves experiment using the MC-MLM loss, on increasing amounts of data, where for each training size we train 3 models and report the mean and std, and report the results in Table 6. Here, in contrast to the trends shown in Liu et al. (2020), we observe generally worse results using the MC-MLM loss. One source of difference is that Liu et al. (2020) repeated the experiments much more times while performing a grid search over different hyperparameters, while we used the same default hyperparameters for all experiments. Another source of difference is the different training and evaluation splits used in our studies.

We conclude that nevertheless, the trends remain the same, and the slopes of both methods are slow to increase, and thus strengthens our claims about the limited usefulness of training data for WS.

¹²<https://commoncrawl.org/>

¹³<http://index.commoncrawl.org/CC-MAIN-2019-09/>

# Training	BERT		RoBERTa		ALBERT	
	Single	Group	Single	Group	Single	Group
0	52.99 (0.00)	8.67 (0.00)	56.39 (0.00)	16.61 (0.00)	55.55 (0.00)	17.23 (0.00)
100	54.39 (1.59)	12.28 (2.39)	55.46 (0.18)	17.61 (1.46)	56.14 (1.12)	17.54 (3.73)
500	51.32 (0.37)	10.53 (2.14)	55.63 (1.67)	25.00 (3.27)	61.97 (1.37)	34.15 (1.74)
1,000	51.75 (0.63)	12.28 (0.89)	58.27 (2.03)	35.56 (3.27)	62.85 (0.12)	34.86 (0.25)
2,000	54.93 (0.37)	14.44 (1.33)	57.92 (1.09)	35.21 (3.16)	61.44 (0.75)	34.51 (0.50)
4,000	52.46 (0.71)	16.55 (2.00)	61.09 (1.84)	37.32 (2.67)	64.08 (2.49)	40.49 (2.32)
6,000	53.87 (1.57)	20.07 (1.53)	59.15 (0.62)	37.32 (1.51)	68.66 (1.12)	49.65 (0.75)
8,000	53.69 (1.17)	22.89 (1.08)	62.15 (0.98)	39.44 (1.27)	68.13 (2.74)	50.70 (3.73)
10,000	53.87 (0.51)	25.00 (1.61)	63.56 (1.24)	45.77 (2.21)	70.42 (1.49)	53.17 (0.50)
12,000	50.17 (1.50)	23.94 (2.46)	64.26 (1.07)	45.42 (3.27)	69.54 (0.51)	52.46 (0.50)
14,000	52.82 (2.00)	27.11 (3.86)	63.38 (1.25)	44.72 (2.99)	67.61 (0.97)	53.17 (1.81)
16,000	53.69 (0.67)	27.11 (0.89)	61.09 (0.57)	41.67 (1.77)	70.77 (0.75)	55.28 (1.23)

Table 6: Effect of the training data size on different models performance. We report results on BERT, RoBERTa and ALBERT, all with their largest variants.