WinoGrande: An Adversarial Winograd Schema Challenge at Scale

By Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi

Abstract

Commonsense reasoning remains a major challenge in AI, and yet, recent progresses on benchmarks may seem to suggest otherwise. In particular, the recent neural language models have reported above 90% accuracy on the Winograd Schema Challenge (WSC),[22] a commonsense benchmark originally designed to be unsolvable for statistical models that rely simply on word associations. This raises an important question—whether these models have truly acquired robust commonsense capabilities or they rely on spurious biases in the dataset that lead to an overestimation of the true capabilities of machine commonsense.

To investigate this question, we introduce WinoGrande, a large-scale dataset of 44k problems, inspired by the original WSC, but adjusted to improve both the scale and the hardness of the dataset. The key steps of the dataset construction consist of (1) large-scale crowdsourcing, followed by (2) systematic bias reduction using a novel AFLITE algorithm that generalizes human-detectable *word associations* to machine-detectable *embedding associations*. Our experiments demonstrate that state-of-the-art models achieve considerably lower accuracy (59.4%–79.1%) on WINOGRANDE compared to humans (94%), confirming that the high performance on the original WSC was inflated by spurious biases in the dataset.

Furthermore, we report new state-of-the-art results on five related benchmarks with emphasis on their dual implications. On the one hand, they demonstrate the effectiveness of WINOGRANDE when used as a resource for transfer learning. On the other hand, the high performance on all these benchmarks suggests the extent to which spurious biases are prevalent in all such datasets, which motivates further research on algorithmic bias reduction.

1. INTRODUCTION

Commonsense reasoning has been a long-standing open research question in AI.[5] The Winograd Schema Challenge (WSC),[22] proposed as an alternative to the Turing Test,[39] has been regarded as a prototypical benchmark to test commonsense capabilities in AI. WSC are designed to be pronoun resolution problems (see examples in Table 1) that are trivial for humans but hard for machines that merely rely on statistical patterns such as word associations without true commonsense understanding. One of the difficulties in commonsense reasoning comes from "reporting bias" in language[15]; commonsense knowledge is often too obvious for people to explicitly state in text, which can confuse the models that rely on statistical patterns in language.

However, recent advances in neural language models have saturated most major benchmarks, such as a variant of WSC dataset where the models now achieve around 90% accuracy. This raises a curious question:

Have neural language models successfully acquired commonsense or are we overestimating the true capabilities of machine commonsense?

This question about the potential overestimation leads to another crucial question regarding potential unwanted biases that the large-scale neural language models might be exploiting, essentially solving the problems *right*, but for *wrong* reasons. Indeed, although WSC questions are carefully crafted by experts, recent studies have shown that they are nevertheless prone to incidental biases. Trichelair et al.[36] have reported *word-association* (13.5% of the cases, see Table 1 for examples) as well as other types of *dataset-specific* biases. Although such biases and annotation artifacts are not apparent for individual instances, they get introduced in the dataset as problems as authors subconsciously repeat similar problem-crafting strategies.

To investigate this question about the true estimation of the machine commonsense capabilities, we introduce **Wino-Grande**, a new dataset with 44k problems that are inspired by the original design of WSC, but modified to improve both the scale and hardness of the problems. The key steps in WINOGRANDE construction consist of (1) a carefully designed crowdsourcing procedure, followed by (2) a novel algorithm AFLITE that generalizes human-detectable biases based on *word* occurrences to machine-detectable biases based on *embedding* occurrences. The key motivation of our approach is that it is difficult for humans to write problems without accidentally inserting unwanted biases.

Although humans find WINOGRANDE problems trivial with 94% accuracy, the best state-of-the-art results, such as those from RoBERTa,[25] are considerably lower (59.4%–79.1%) depending on the amount of training data provided (from 800 to 41k instances). Furthermore, we also demonstrate that WINOGRANDE provides transfer learning to other existing WSC and related benchmarks, achieving new state-of-the-art (SOTA) performances.

Although the improvements of SOTA over multiple challenging benchmarks are exciting, we cautiously note that these positive results must be taken with a grain of salt. The result might also indicate the extent to which spurious effects are prevalent in existing datasets, which runs the risk of overestimating the true capabilities of

Table 1. WSC problems are constructed as pairs (called *twin*) of nearly identical questions with two answer choices.

		Twin sentences	Options (answer)
✔ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy**/suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy/**suitcase**
✔ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann**/Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann/**Mary**
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree**/roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree/**roof**
✗ (4)	a	The lions ate the zebras because **they** are *predators*.	**lions**/zebras
	b	The lions ate the zebras because they are meaty.	lions/**zebras**

The questions include a *trigger word* that flips the correct answer choice between the questions. Examples (1)–(3) are drawn from WSC[22] and (4) from DPR.[31] Examples marked with ✗ have language-based bias that current language models can easily detect. Example (4) is undesirable because the word "predators" is more often associated with the word "lions," compared to "zebras."

machine intelligence on commonsense reasoning. More generally, human-crafted problems and tasks (regardless of whether they are crowd-sourced or by experts) contain annotation artifacts in many cases, and algorithmic bias reduction such as AFLITE is essential to mitigate such dataset-specific bias.

Our work suggests a new perspective for measuring progress in AI. Instead of constructing *static* benchmark datasets and asking the community to work on them for years, we propose the use of *dynamic* datasets that evolve together with the state-of-the-art models.

2. CROWDSOURCING WINOGRANDE AT SCALE
WSC problems have been considered challenging to craft by crowdsourcing due to the structural constraints of twins and the requirement of linguistic knowledge (Table 1). Nevertheless, we present an effective approach to creating a large-scale dataset (WINOGRANDE) of WSC problems while maintaining its original properties—that is, trivial for humans but hard for AI systems. Our approach consists of a carefully designed crowdsourcing task followed by a novel adversarial filtering algorithm (§3) that systematically removes biases in the data.

Enhancing crowd creativity. Creating twin sentences from scratch puts a high cognitive load on crowd workers who thereby subconsciously resort to writing pairs that are lexically and stylistically repetitive. To encourage creativity and reduce their cognitive load, we employed *creativity from constraints*[35]— a psychological notion which suggests that appropriate constraints can help structure and drive creativity. In practice, crowd workers are primed by a randomly chosen topic as a suggestive context (details here), although they are asked to follow precise guidelines on the structure of the curated data.

Crowdsourcing task. We collect WINOGRANDE problems via crowdsourcing on Amazon Mechanical Turk (AMT). Workers are asked to write twins sentences (as shown in Table 1) that meet the requirements for WSC problems (e.g., avoiding word association, nonzero but small edit distance). To avoid repeating the same topics, workers were instructed to randomly pick an *anchor* word(s) from a randomly assigned WikiHow article and to ensure that the twin sentences contain the *anchor* word. The *anchor* word does not have to be a *trigger* word, but we ensured that it is not a function word

such as *it, the,* and *of.* In our pilot experiments, we found that this constraint drastically improves the worker's creativity and diversity of topics. Additionally, workers were instructed to keep twin sentence length in between 15 and 30 words although maintaining at least 70% word overlap between a pair of twins. Following the original WSC problems, we aimed to collect twins in two different domains—(i) social commonsense: a situation involving two same gender people with contrasting attributes, emotions, social roles, etc., and (ii) physical commonsense: a context involving two physical objects with contrasting properties, usage, locations, etc. In total, we collected 77k questions (i.e., 38k twins).

Data validation. We validated each collected question through a distinct set of three crowd workers. Following a rigorous process, a question is deemed valid if (1) the majority of the three workers chooses the correct answer option, (2) they agree that the two answer options are unambiguous (one option is clearly more plausible than the other), and (3) the question cannot be answered simply by word association in which the local context around the target pronoun is given (e.g., "because **it** was going so fast." (**race car**/school bus)). As a result, 68% of the questions (53k) were deemed valid and we discarded the invalid questions.

Although our crowdsourcing procedure addresses some amount of instance-level biases such as word association, it is still possible that the constructed dataset has *dataset-specific* biases, especially after it has been scaled up. To address this challenge, we propose a method for systematic bias reduction.

3. ALGORITHMIC DATA BIAS REDUCTION
Several recent studies[16, 29, 38, 27, 12] have reported the presence of *annotation artifacts* in large-scale datasets. Annotation artifacts are unintentional patterns in the data that leak information about the target label in an undesired way. State-of-the-art neural models are highly effective at exploiting such artifacts to solve problems *correctly*, but for *incorrect* reasons. To tackle this persistent challenge with dataset biases, we propose AFLITE—a novel algorithm that can systematically reduce biases using the state-of-the-art contextual representation of words.

Lightweight adversarial filtering. Our approach builds upon the adversarial filtering (AF) algorithm proposed by Zellers et al.,[41] but makes two key improvements: (1) AFLITE is much more broadly applicable (by not requiring over generation of data instances) and (2) it is considerably more lightweight (not requiring retraining a model at each iteration of AF). Overgenerating machine text from a language model to use in test instances runs the risk of distributional bias where a discriminator can learn to distinguish between machine generated instances and human-generated ones. In addition, AF depends on training a model at each iteration, which comes at extremely high computation cost when being adversarial to a model such as BERT.[7]

Instead of manually identified lexical features, we adopt a dense representation of instances using their *pre-computed* neural network embeddings. In this work, we use RoBERTa[25] fine-tuned on a small subset of the dataset. Concretely, we use 6k instances (5k for training and 1k for validation) from the dataset (containing 53k instances in total) to fine-tune RoBERTa (referred to as RoBERTa$_{embed}$). We use RoBERTa$_{embed}$ to pre-compute the embeddings for the rest of the instances (47k) as the input for AFLITE. We discard the 6k instances from the final dataset.

Next, we use an ensemble of linear classifiers (logistic regressions) trained on random subsets of the data to determine whether the representation used in RoBERTa$_{embed}$ is strongly indicative of the correct answer option. If so, we discard the corresponding instances and proceed iteratively.

Figure 1 provides an illustration of AFLITE algorithm. The algorithm takes as input the *pre-computed* embeddings and labels, along with the size n of the ensemble, the training size m for the classifiers in the ensemble, the size of the filtering cutoff, and the filtering threshold τ. At each filtering phase, we train n linear classifiers on different random partitions of the data and we collect their predictions on their corresponding validation set. For each instance, we compute its *score* as the ratio of correct predictions over the total number of predictions. We rank the instances according to their score and remove the top-k instances whose score is above threshold τ. We repeat this process until we remove fewer than k instances in a filtering phase or there are fewer than m remaining instances. When applying AFLITE to WINOGRANDE, we set $m = 10,000$, $n = 64$, $k = 500$, and $\tau = 0.75$.

This approach is also reminiscent of recent work in NLP on adversarial learning.[3, 1, 9] Belinkov et al.[1] proposed an adversarial removal technique for NLI, which encourages models to learn representations that are free of hypothesis-only biases. When proposing a new benchmark, however, we cannot enforce that any future model will purposefully avoid learning spurious correlations in the data. In addition, although the hypothesis-only bias is an insightful bias in NLI, we make no assumption about the possible sources of

The AfLite algorithm is published with further development.[20]
AfLite is designed for filtering instances so that the resulting dataset is less biased, whereas the original AF algorithm[41] is designed for "generating and modifying" individual instances, such as by creating better distractors. AfLite and AF are therefore different in their goals and hence difficult to compare directly.

Figure 1. Illustration of the AfLite algorithm. It takes as input the pre-computed representations of each instance (e.g., BERT embeddings). An ensemble of linear classifiers are trained on different random partitions of the data and used to compute the predictability score for each instance. The algorithm filters out the instances with the highest scores and proceeds iteratively to the next filtering phase.
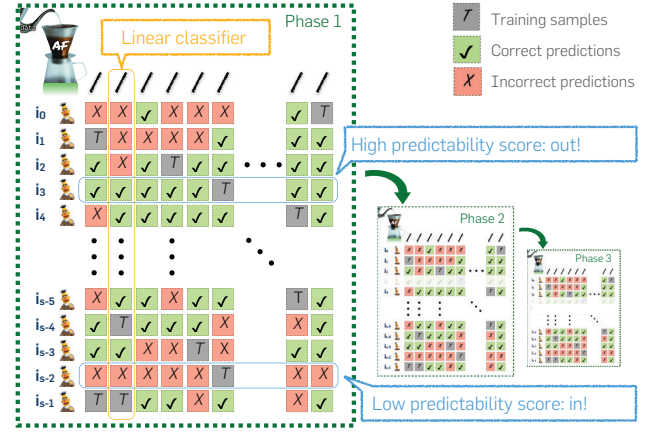
bias in WINOGRANDE. Instead, we adopt a more proactive form of bias reduction by relying on the state-of-the-art (statistical) methods to uncover undesirable dataset shortcuts.

Assessment of AfLite. We assess the impact of AFLITE relative to two baselines: random data reduction and pointwise mutual information (PMI) filtering. In random data reduction, we randomly subsample the dataset to evaluate how a decrease in dataset size affects the bias. In PMI filtering, we compute the difference (f) of PMIs for each twin (t) as follows:

$$f(t_1, t_2) = \sum_{w \in t_1} \mathrm{PMI}(y = 1; w) - \sum_{w \in t_2} \mathrm{PMI}(y = 1; w).$$

Technically, we first pre-computed PMI between a word and the label $y = 1$ for each word in the dataset, following a method proposed by Gururangan et al.[16] The sum of PMI value of each token in a given sentence indicates the likelihood of the label $y = 1$ for the sentence. We only retain the twins that have a small difference in their PMI values as it corresponds to the twins that are hard to discriminate.

Figure 2 plots RoBERTa pre-computed embeddings whose dimension is reduced to 2D (*top*) and 1D (*bottom*) using Principal Component Analysis (PCA). We observe that WINO-GRANDE$_{all}$ and the two baselines exhibit distinct components between the two correct answer options (i.e., $y \in 1, 2$), whereas such distinction becomes less salient in WINO-GRANDE$_{debiased}$, which implies that AFLITE successfully reduces the spurious correlation in the dataset (between instances and labels). To quantify the effect, we compute the KL divergence between the samples with answer options. We find that the random data reduction does not reduce the KL divergence ($2.53 \rightarrow 2.51$). It is interesting to see that PMI-filtering marginally reduces the KL divergence ($\rightarrow 2.42$), although the principal component analysis on the PMI-filtered subset still leads to a significant separation between the labels. On the other hand, in WINOGRANDE$_{debiased}$, AFLITE reduces the KL divergence

Figure 2. The effect of debiasing by AfLite. RoBERTa pre-computed embeddings (applied PCA for dimension reduction) are shown in two-dimensional space (*top row*) and histograms regarding d_1 (*bottom row*) with the bin size being 100. Data points are colored depending on the label (i.e., the answer y is option 1 (blue) or 2 (red)). In the histograms, we show the KL-divergence between $p(d_1, y=1)$ and $q(d_1, y=2)$.
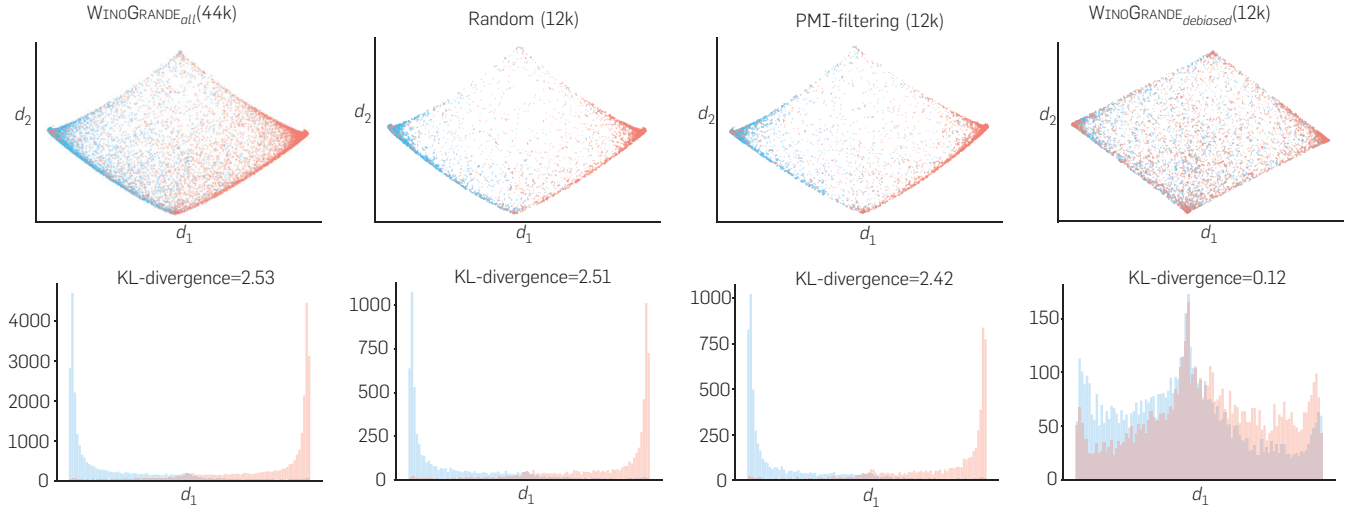
WINOGRANDE$_{all}$(44k)   Random (12k)   PMI-filtering (12k)   WINOGRANDE$_{debiased}$(12k)

KL-divergence=2.53   KL-divergence=2.51   KL-divergence=2.42   KL-divergence=0.12

Table 2. Examples that have *dataset-specific* bias detected by AfLite (marked with ✗).

	Twin sentences	Options (answer)
✗	The monkey loved to play with the balls but ignored the blocks because he found **them** *exciting*. The monkey loved to play with the balls but ignored the blocks because he found **them** *dull*.	**balls**/blocks balls/**blocks**
✗	William could only climb beginner walls while Jason climbed advanced ones because **he** was very *weak*. William could only climb beginner walls while Jason climbed advanced ones because **he** was very *strong*.	**William**/Jason William/**Jason**
✔	Robert woke up at 9:00 am while Samuel woke up at 6:00 am, so **he** had *less* time to get ready for school. Robert woke up at 9:00 am while Samuel woke up at 6:00 am, so **he** had *more* time to get ready for school.	**Robert**/Samuel Robert/**Samuel**
✔	The child was screaming after the baby bottle and toy fell. Since the child was *hungry*, **it** stopped his crying. The child was screaming after the baby bottle and toy fell. Since the child was *full*, **it** stopped his crying.	**baby bottle**/toy baby bottle/**toy**
	The words that include (dataset-specific) polarity bias (§3) are highlighted (positive and negative). For comparison, we show examples selected from WinoGrande$_{debiased}$ (marked with ✔).	

dramatically ($\rightarrow$ 0.12), which suggests that this debiased dataset should be challenging for statistical models that solely rely on spurious correlation.

What bias has been actually detected by AfLite? Is the bias really spurious and undesirable according to the original WSC's goal? Table 2 presents examples that AFLITE has detected as a dataset-specific bias. We see a structural pattern in the first two twins, where the sentiment between the answer option and the target pronoun is highly correlated. In other words, these problems can be easily answered by simply exploiting the pattern of the polarity (positive or negative). Importantly, this dataset-specific bias is structural rather than at the token level, contrasting with the biases that have been identified in the NLI literature,[16, 29] and it is hard to detect these biases using heuristics such as lexical PMI-filtering. Instead of depending on such heuristics, AFLITE is able to detect the samples that potentially have such biases algorithmically.

After applying the AFLITE algorithm, we obtain a *debiased* dataset of 12,282 instances split into training (9,248), development (1,267), and test (1,767) sets. We also release 31k

problems that are filtered out by AFLITE for additional training set (§4) and resource (§5), resulting in a total number of problems in WINOGRANDE$_{all}$ to be 43,432 (40,398 for training, 1,267 for development, and 1,767 for test).

WinoGrande versus the Original WSC. Although WINO-GRANDE is inspired by the original WSC, we make a few design choices that deviate from the original design guidelines of WSC in order to scale up the dataset considerably while ensuring the hardness of the dataset.

First, WINOGRANDE is formatted as a fill-in-the-blank problem where the blank corresponds to the mention of one of the two names in the context, following the same modification made by other recent WSC variants such as Trinh and Le.[37] By contrast, the original WSC explicitly places a pronoun (instead of a blank). From the modeling stand point, the use of blanks instead of explicit pronouns do not make the problem any easier.

Second, although we originally collected all problems in twins, the final questions in the filtered WINOGRANDE$_{debiased}$ are not always twins, because it is possible that AFLITE filters out only one of the twin sentences. In WINOGRANDE$_{debiased}$,

about 1/3 of questions are not twins. We also release WINO-GRANDE$_{all}$ (training set) that all consists of twins.

Third, unlike the original WSC problems that were composed by just a few linguistics experts, WINOGRANDE is authored by crowdworkers. Thus, the language used in WINOGRANDE reflects the more diverse and noisy language used by crowds. Importantly, laymen still find WINOGRANDE problems easy to solve, with 94% accuracy (§4).

4. EXPERIMENTAL RESULTS

4.1. Baseline models
We evaluate the WINOGRANDE$_{debiased}$ (dev and test) on the methods/models that have been effective on the original WSC.

Wino knowledge hunting. Wino Knowledge Hunting (WKH) by Emami et al.[10] is based on an information retrieval approach, where the sentence is parsed into a set of queries and then the model looks for evidence for each answer candidate from the search result snippets.

Ensemble neural LMs. Trinh and Le[37] is one of the first attempts to apply a neural language model, which is pretrained on a very large corpora (such as LM-1-Billion, CommonCrawl, SQuAD, and Gutenberg Books). In this approach, the task is treated as fill-in-the-blank question with binary choice. The target pronoun in the sentence is replaced by each answer candidate, and the neural language model provides the likelihood of the two resulting sentences. This simple yet effective approach outperforms previous IR-based methods.

BERT. BERT[7] is another pre-trained neural language model that has bidirectional paths and consecutive sentence representations in hidden layers. We finetune BERT with splitting the input sentence into context and option using the candidate answer as delimiter. The input format becomes [CLS] context [SEP] option [SEP]; for example, *The trophy doesn't fit into the brown suitcase because the _____ [SEP] is too large. [SEP]* (The blank _____ is filled with either option 1 or 2), and the [CLS] token embedding is used to classify which answer option is correct. We used grid-search for hyper-parameter tuning: learning rate $\{1e-5, 3e-5, 5e-5\}$, number of epochs $\{3, 4, 5, 8\}$, and batch-size $\{8, 16\}$ with three different random seeds.

RoBERTa. RoBERTa[25] is an improved variant of BERT that adds more training data with larger batch sizes and training time, as well as other refinements such as dynamic masking. RoBERTa performs consistently better than BERT across many benchmark datasets.

Word association baseline. Using BERT and RoBERTa, we also run the word association baseline (*local-context-only*) to check if the dataset can be solved by language-based bias. In this baseline, the model is trained with only local contexts ($w_{t-2:EOS}$) surrounding the blank to be filled (w_t) (e.g., because the _____ [SEP] is too large. [SEP]). This is analogous to the *hypothesis-only* baseline in NLI,[29] where the task (dataset) does not require the full context to achieve high performance.

https://github.com/tensorflow/models/tree/master/research/lm_commonsense

Table 3. Performance of several baseline systems on WinoGrande$_{debiased}$ (dev and test).

Methods	Dev acc. (%)	Test acc. (%)
WKH	49.4	49.6
Ensemble LMs	53.0	50.9
BERT	65.8	64.9
RoBERTa	**79.3**	**79.1**
BERT (local context)	52.5	51.9
RoBERTa (local context)	52.1	50.0
BERT-DPR★	50.2	51.0
RoBERTa-DPR★	59.4	58.9
Human Perf.	94.1	94.0

The star (★) denotes that it is zero-shot setting (e.g., BERT-DPR★ is a BERT model fine-tuned with the DPR dataset and evaluated on WinoGrande$_{debiased}$).

Table 4. Performance of RoBERTa with different training sizes.

Training size	Dev acc. (%)	Test acc.(%)
XS (160)	51.5	50.4
S (640)	58.6	58.6
M (2,558)	66.9	67.6
L (10,234)	75.8	74.7
XL (40,398)	79.3	79.1

Fine-tuning on DPR dataset. Definite Pronoun Resolution (DPR) Dataset, collected by Rahman and Ng,[31] consists of 1,886 WSC style problems written by 30 undergraduate students. Kocijan et al.[19] have recently shown that BERT finetuned with DPR boosts the performance on WCS (72.2% accuracy). As additional baselines, we finetune BERT and RoBERTa with DPR and evaluate on WINO-GRANDE. This allows us to compare the difficulty of WSC and WINOGRANDE empirically.

Human evaluation. In addition to the methods described above, we compute human performance as the majority vote of three crowd workers for each question.
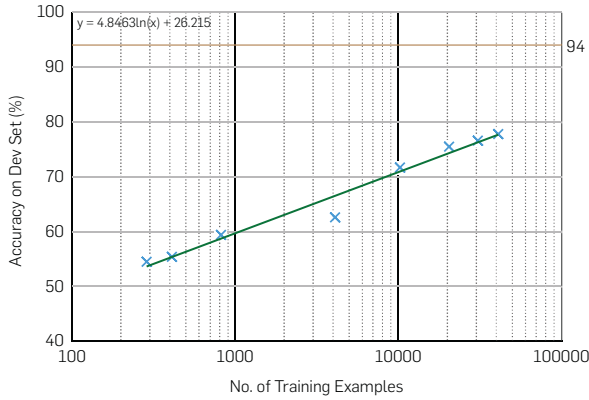
4.2. Results
Table 3 shows the results. Two baselines, WKH and Ensemble LMs, only achieve chance-level performance (50%). The best model, RoBERTa, achieves 79.1% test-set accuracy, whereas human performance achieves 94.0%, indicating that the WINOGRANDE$_{debiased}$ is still easy for humans to answer as desired. Regarding the word association (i.e., local context) baselines, both BERT and RoBERTa achieve close to chance-level performance, illustrating that most WINOGRANDE$_{debiased}$ problems cannot be answered by local context only. Finally, BERT and RoBERTa finetuned with DPR achieve chance-level to below 60% accuracy, which contrast with the performance boosts on WSC (72% by BERT (Kocijan et al.[19]) and 83% in RoBERTa) and other existing WSC-style problems (as shown in §5.3). This indicates that WINOGRANDE$_{debiased}$ consists of more challenging problems than WSC and existing variants.

When we use the debiased training set (9248), both BERT and RoBERTa showed only chance level performance.

Figure 3. Learning curve on the dev set of Wino-Grande. Each point on the plot is the best performance for a given number of randomly selected training examples, computed over 10 random seeds.

Table 5. Statistics on WSC and related datasets (§5.1).

Dataset	#Probs	Avg Len	#Vocab
WSC	273	19.1	919
PDP	80	39.5	594
SuperGLUE-WSC	804	28.4	1711
DPR	1886	15.9	4127
KnowRef	1269	19.3	5310
COPA	1000	13.3	3369
Winogender	720	15.6	523
WINOGRANDE$_{debiased}$	12,282	21.1	11,408
WINOGRANDE$_{all}$	43,432	20.6	16,469

Learning curve. In order to see the effect of training size, Table 4 shows the performance by RoBERTa trained on different training sizes from 160k to 40k questions. Figure 3 shows the learning curve of the best model, RoBERTa, on the WINOGRANDE$_{debiased}$ dev set. RoBERTa's performance ranges from 59% to 79% when the size of training data is varied from 800 (2% of the training data) to 41K (100% of the training data) instances. To achieve human-level performance, the current state-of-the-art models would need over 118K training instances.

Importantly, the lower end of the available training data (∼800) in the learning curve roughly matches the size of the training data made available in previous variants of WSC (see Table 5). For most of these datasets, state of the art already reaches around 90% (§5). By contrast, when we control for the training set size in WINOGRANDE, RoBERTa's performance is considerably lower (59%), demonstrating that our dataset construction method is able to compose WSC problems that are collectively considerably harder than previous datasets.

5. TRANSFER LEARNING FROM WINOGRANDE
WINOGRANDE contains a large number of WSC style questions. In addition to serving as a benchmark dataset, we use WINOGRANDE as a resource—we apply transfer learning by first fine-tuning a model on our dataset and evaluating its performance on related datasets: WSC, PDP, SuperGLUE-WSC, DPR, KnowRef, KnowRef, and Winogender. We establish the state-of-the-art results across several of these existing benchmark datasets.

5.1. Existing WSC and related datasets
We briefly describe existing WSC variants and other related datasets. Table 5 provides their summary statistics.

WSC.[22] This is the original Winograd Schema Challenge dataset, which consists of 273 problems. The problems are manually crafted by the authors to avoid word association bias as much as possible, although Trichelair et al.[36] later report that 13.5% of the questions may still have word-association bias.

PDP.[26] Pronoun Disambiguation Problems (PDP) dataset is closely related to the original WSC, and used in the 2016 running of the Winograd Schema Challenge. The dataset consists of 80 pronoun disambiguation problems. It is formulated as a multiple choice task, in which a pronoun must be resolved to one of up to 5 (but mostly binary) possible antecedents.

SuperGLUE-WSC.[40] SuperGLUE contains multiple datasets such as a modified version of WSC, which we will refer to as SuperGLUE-WSC. This dataset aggregates the original WSC, PDP and additional PDP-style examples, and recasts them into True/False binary problems (e.g., "Pete envies **Martin** because *he* is very successful." Q: Does *he* refer to **Martin**? A: True). The number of problems are roughly doubled from WSC and PDP, although the size is still relatively small (804 in total). We converted WinoGrande to the True/False binary problems.

DPR.[31] Definite Pronoun Resolution Dataset (DPR) introduces 1,886 additional WSC problems authored by 30 undergraduate students. Trichelair et al.[36] point out that this dataset is overall less challenging than the original WSC due to an increased level of language-based or dataset-specific biases. We split the original training set (1,332) into training (1,200) and development (122) sets, DPR does not have an official split for it.

KnowRef.[11] KnowRef provides over 8k WSC-style coreference resolution problems that are extracted and filtered with heuristic rules from 100 million web sentences (Reddit, Wikipedia, and OpenSubtitles). We report results on the publicly available *test* set (1.2k problems).

COPA.[32] This dataset introduces 1000 problems that aim to test commonsense reasoning focusing on script knowledge, formulated as a binary choice about *causes* and *effects* of given premises. Because COPA does not provide a training set, we split the original development set (500) into training (400) and development (100) sets in the same way as SuperGLUE-COPA.[40]

Winogender.[33] This dataset introduces 720 problems focusing on pronouns resolution with respect to people, with distinct goal of measuring gender bias in coreference resolution systems.

Since the original publication of this paper, there have been several updates with higher performance such as Lin et al.[23] and Khashabi et al.[18] that rely on similar models with even larger parameters and data sources, implying that the models detect annotation artifacts better than RoBERTa. This indicates that we need *dynamic* datasets that evolve together with the evolving state-of-the-art algorithms.

5.2. Experimental setup

Our model is based on RoBERTa finetuned with WINO-GRANDE (train and dev sets). To compare different corpora used as a resource, we also finetune RoBERTa on DPR (train and test sets). For hyper parameter search, we use the same grid search strategy as in §4.

Additional human evaluation. We also report human performance for WSC, PDP, and DPR to calibrate the quality of our crowd worker pool as well as support previous findings. To our knowledge, this is the first work to report human performance on the DPR dataset.

5.3. Experimental results

Tables 6 and 7 show the results of applying transfer learning from WINOGRANDE to other WSC variants. Overall, RoBERTa fine-tuned on WINOGRANDE helps improve the accuracy on all the related tasks (Table 6), and performs consistently better than when it is fine-tuned on DPR.

Although improvements on some related datasets (particularly WSC, PDP, and DPR) might seem expected, the significant improvement on COPA is not so. The COPA task—identifying causes and effects—is very different from that in WINOGRANDE. This significant improvement on an unrelated task indicates that WINOGRANDE can serve as a resource for commonsense knowledge transfer.

Important implications. We consider that although these positive results over multiple challenging benchmarks are highly encouraging, they may need to be taken with a grain of salt. In particular, these results might also indicate the extent to which spurious dataset biases are prevalent in existing datasets, which runs the risk of overestimating the true capabilities of machine intelligence on commonsense reasoning.

Our results and analysis indicate the importance of continued research on debiasing benchmarks and the increasing need for algorithmic approaches for systematic bias reduction, which allows for the benchmarks to evolve together with evolving state of the art. We leave it as a future research question to further investigate how much of our improvements are due to dataset biases of the existing benchmarks as opposed to true strides in improving commonsense intelligence.

5.4. Diagnostics for gender bias

Winogender is designed as diagnostics for checking whether a model (and/or training corpora) suffers from gender bias. The bias is measured by the difference in accuracy between the cases where the pronoun gender matches the occupation's majority gender (called "non-gotcha") or not ("gotcha"). Formally, it is computed as follows:

$$\Delta F = \text{Acc}_{(\text{Female,Non-gotcha})} - \text{Acc}_{(\text{Female,Gotcha})}$$

$$\Delta M = \text{Acc}_{(\text{Male,Non-gotcha})} - \text{Acc}_{(\text{Male,Gotcha})}$$

for female and male cases, respectively.

Large values of ΔF or ΔM indicate that the model is highly gender-biased, whereas $|\Delta F| = |\Delta M| = 0$ (along with high accuracy) is the ideal scenario. In addition, if ΔF or ΔM is largely *negative*, it implies that the model is biased in the other way around.

The result of the gender-bias diagnostics is shown in Table 7. Although we find that the RoBERTa models

Table 6. Accuracy (%) on existing WSC-related tasks (test set).

WSC[22]	
Liu et al.[24]	52.8
WKH[10]	57.1
Ensemble LMs[37]	63.8
GPT2[30]	70.7
BERT-DPR★[19]	72.2
HNN[17]	75.1[†]
RoBERTa-DPR★ (This work)	83.1
RoBERTa-WinoGrande★ (This work)	**90.1**
Humans[2]	92.1
Humans (This work)	96.5
PDP[26]	
Liu et al.[24]	61.7
Trinh and Le[37]	70.0
RoBERTa-DPR★ (This work)	86.3
RoBERTa-WinoGrande★ (This work)	87.5
HNN[17]	**90.0[†]**
Humans[6]	90.9
Humans (This work)	92.5
SuperGLUE-WSC[40]	
Majority baseline	65.1
RoBERTa-DPR-ft (This work)	83.6
RoBERTa-WinoGrande-ft (This work)	85.6
RoBERTa-ensemble[25]	**89.0**
Humans[40]	100
DPR[31]	
Rahman and Ng[31]	73.0
Peng et al.[28]	76.4
BERT-WinoGrande★ (This work)	84.9
RoBERTa-ft (This work)	91.7
RoBERTa-WinoGrande★ (This work)	92.5
RoBERTa-WinoGrande-ft (This work)	**93.1**
Humans (This work)	95.2
KnowRef[11]	
Emami et al.[11]	65.0
RoBERTa-DPR★ (This work)	84.2
RoBERTa-WinoGrande★ (This work)	**85.6**
Humans[11]	92.0
COPA[32]	
Gordon et al.[14]	65.4
Sasaki et al.[34]	76.4
RoBERTa-WinoGrande★ (This work)	84.4
RoBERTa-ft (This work)	86.4[‡]
RoBERTa-WinoGrande-ft (This work)	**90.6**
Humans[13]	99.0

The star (★) denotes that it is zero-shot setting. "-ft" indicates *fine-tuning* on the targeted dataset (train and dev). RoBERTa-X-ft denotes sequential fine-tuning with dataset X followed by the targeted dataset. The dagger ([†]) indicates that the evaluation data is not exactly the same from ours. The double dagger ([‡]) denotes that we could not reproduce he same number as in SuperGLUE leaderboard.[40]

fine-tuned on WINOGRANDE and DPR both demonstrate very high-accuracy, the gender gap in RoBERTa-WinoGrande is smaller than RoBERTa-DPR.

6. CONCLUSION

We introduce WINOGRANDE, a new collection of 44k WSC-inspired problems that is significantly larger than existing variants of the WSC dataset. To create a dataset that is robust

Table 7. Accuracy (%) and gender bias on Winogender dataset.

Winogender[33]					
	Gotcha	Female	Male	|ΔF|	|ΔM|
Rule	No	38.3	51.7	28.3	14.2
	Yes	10.0	37.5		
Stats	No	50.8	61.7	5.0	21.7
	Yes	45.8	40.0		
Neural	No	50.8	49.2	14.1	2.5
	Yes	36.7	46.7		
RoBERTa-DPR (This work)	No	98.3	96.7	1.6	0.9
	Yes	96.7	95.8		
RoBERTa-WG (This work)	No	97.5	96.7	0.8	0.8
	Yes	96.7	97.5		

"Gotcha" indicates whether the target gender pronoun (e.g., she) is minority in the correct answer option (e.g., doctor). |ΔF| and |ΔM| show the system performance gap between "Gotcha" and "non-Gotcha" for each gender (lower the better). The first three baselines are adopted from Rudinger et al.[33]; Rule is Lee et al.,[21] Stats is Durrett and Klein,[8] and Neural is Clark and Manning.[4]

against spurious dataset-specific bias, we also present AFLITE—a novel lightweight adversarial filtering algorithm for systematic bias reduction. The resulting dataset is considerably more challenging for existing state-of-the-art models while still being trivially easy for humans. In addition, using WINOGRANDE as a resource, we demonstrate effective transfer learning and achieve state-of-the-art results on several related benchmarks.

In parallel, we also emphasize the potential risk of overestimating the performance of the state-of-the-art methods on the existing commonsense benchmarks; these models might be solving the problems *right* for the *wrong* reasons, by relying on spurious statistical patterns (annotation artifacts).

Our work suggests a new perspective for designing benchmarks for measuring progress in AI. Unlike past decades where the community constructed a *static* benchmark dataset to work on for many years to come, we now need AI algorithms to compose challenges that are hard enough for AI, which requires *dynamic* datasets that evolve together with the evolving state-of-the-art.

References
1. Belinkov, Y., Poliak, A., Shieber, S., Van Durme, B., Rush, A. On adversarial removal of hypothesis-only bias in natural language inference. *SEM (2019), 256–262.
2. Bender, D. Establishing a human baseline for the winograd schema challenge. *MAICS* (2015), 30–45.
3. Chen, X., Cardie, C. Multinomial adversarial networks for multi-domain text classification. *NAACL* (2018), 1226–1240.
4. Clark, K., Manning, C.D. Deep reinforcement learning for mention-ranking coreference models. *EMNLP* (2016), 2256–2262.
5. Davis, E., Marcus, G. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM 58*, 9 (Aug. 2015), 92–103.
6. Davis, E., Morgenstern, L., Ortiz, C. Human tests of materials for the winograd schema challenge Unpublished manuscript (2016). https://cs.nyu.edu/faculty/davise/papers/WS2016SubjectTests.pdf, 2016.
7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018).
8. Durrett, G., Klein, D. Easy victories and uphill battles in coreference resolution. *EMNLP* (2013), 1971–1982.
9. Elazar, Y., Goldberg, Y. Adversarial removal of demographic attributes from text data. *EMNLP* (2018), 11–21.
10. Emami, A., Trischler, A., Suleman, K., Cheung, J.C.K. A generalized knowledge hunting framework for the winograd schema challenge. *NAACL: SRW* (2018), 25–31.
11. Emami, A., Trichelair, P., Trischler, A., Suleman, K., Schulz, H., Cheung, J.C.K The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. *ACL* (2019), 3952–3961.
12. Geva, M., Goldberg, Y., Berant, J. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. arXiv:1908.07898 (2019).
13. Gordon, A., Kozareva, Z., Roemmele, M. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. *SEM (2012), 394–398.
14. Gordon, A.S., Bejan, C.A., Sagae, K. Commonsense causal reasoning using millions of personal stories. *AAAI* (2011), 1180–1185.
15. Gordon, J., van Durme, B. Reporting bias and knowledge acquisition. *AKBC* (2013), 25–30.
16. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., Smith, N.A. Annotation artifacts in natural language inference data. *NAACL* (2018), 107–112.
17. He, P., Liu, X., Chen, W., Gao, J. A hybrid neural network model for commonsense reasoning. arXiv:1907.11983 (2019).
18. Khashabi, D., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., Hajishirzi, H. Unifiedqa: Crossing format boundaries with a single qa system. arXiv preprint arXiv:2005.00700, (2020).
19. Kocijan, V., Cretu, A.-M., Camburu, O.-M., Yordanov, Y., Lukasiewicz, T. A surprisingly robust trick for the winograd schema challenge. *ACL* (2019), 4837–4842.
20. Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., Choi, Y. Adversarial filters of dataset biases. *ICML* (2020).
21. Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. *CoNLL: Shared Task* (2011).
22. Levesque, H.J., Davis, E., Morgenstern, L. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning* (2011).
23. Lin, S.-C., Yang, J.-H., Nogueira, R., Tsai, M.-F., Wang, C.-J., Lin, J. Ttttackling winogrande schemas. arXiv preprint arXiv:2003.08380 (2020).
24. Liu, Q., Jiang, H., Ling, Z.-H., Zhu, X., Wei, S., Hu, Y. Commonsense knowledge enhanced embeddings for solving pronoun disambiguation problems in winograd schema challenge. arXiv:1611.04146 (2016).
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M.S., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L.S., Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. ArXiv, abs/1907.11692 (2019).
26. Morgenstern, L., Davis, E., Ortiz, C. L. Planning, executing, and evaluating the winograd schema challenge. *AI Magazine 37*, 1 (2016), 50–54.
27. Niven, T., Kao, H.-Y. Probing neural network comprehension of natural language arguments. *ACL* (2019), 4658–4664.
28. Peng, H., Khashabi, D., Roth, D. Solving hard coreference problems. *NAACL* (2015), 809–819.
29. Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., Van Durme, B. Hypothesis only baselines in natural language inference. *SEM (2018), 180–191.
30. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* (2019), 777–789.
31. Rahman, A., Ng, V. Resolving complex cases of definite pronouns: The winograd schema challenge. *EMNLP-CoNLL* (2012).
32. Roemmele, M., Bejan, C.A., Gordon, A.S. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning* (2011).
33. Rudinger, R., Naradowsky, J., Leonard, B., Van Durme, B. Gender bias in coreference resolution. *NAACL* (2018), 15–20.
34. Sasaki, S., Takase, S., Inoue, N., Okazaki, N., Inui, K. Handling multiword expressions in causality estimation. *IWCS* (2017).
35. Stokes, P.D. *Creativity from Constraints: The Psychology of Breakthrough*. Springer Publishing Company, New York, NY, 2005.
36. Trichelair, P., Emami, A., Cheung, J.C.K., Trischler, A., Suleman, K., Diaz, F. On the evaluation of commonsense reasoning in natural language understanding. arXiv:1811.01778 (2018).
37. Trinh, T.H., Le, Q.V. A simple method for commonsense reasoning. arXiv:1806.02847 (2018).
38. Tsuchiya, M. Performance impact caused by hidden bias of training data for recognizing textual entailment. *LREC* (2018), 1506–1511.
39. Turing, A.M. Computing machinery and intelligence. *Mind 59*, 236(1950), 433–460.
40. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R. Superglue: A stickier benchmark for general-purpose language understanding systems. arXiv:1905.00537 (2019).
41. Zellers, R., Bisk, Y., Schwartz, R., Choi, Y. Swag: A large-scale adversarial dataset for grounded commonsense inference. *EMNLP* (2018), 93–104.

Keisuke Sakaguchi, Ronan Le Bras, and Chandra Bhagavatula ([keisukes, ronanlb, chandrab]@allenai.org), Allen Institute for AI, Seattle, WA, USA.

Yejin Choi ([yejin]@cs.washington.edu), University of Washington & Allen Institute for AI, Seattle, WA, USA.