WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale

Keisuke Sakaguchi*, **Ronan Le Bras***, **Chandra Bhagavatula***, **Yejin Choi***†
*Allen Institute for Artificial Intelligence
†University of Washington

Abstract

The Winograd Schema Challenge (WSC), proposed by Levesque et al. (2011) as an alternative to the Turing Test, was originally designed as a pronoun resolution problem that cannot be solved based on statistical patterns in large text corpora. However, recent studies suggest that current WSC datasets, even when composed carefully by experts, are still prone to such biases that statistical methods can exploit. We introduce WINOGRANDE, a new collection of WSC problems that are adversarially constructed to be robust against spurious statistical biases. While the original WSC dataset provided only 273 instances, WINOGRANDE includes 43,985 instances, half of which are determined as adversarial. Key to our approach is a novel adversarial filtering algorithm AFLITE for systematic bias reduction, combined with a careful crowdsourcing design. Despite the significant increase in training data, the performance of existing state-of-the-art methods remains modest (61.6%) and contrasts with high human performance (90.8%) for the binary questions. In addition, WINOGRANDE allows us to use transfer learning for achieving new state-of-the-art results on the original WSC and related datasets. Finally, we discuss how biases lead to overestimating the true capabilities of machine commonsense.

1 Introduction

Commonsense reasoning capability is one of the key differences between human intelligence and modern AI. Most successful modern AI systems rely primarily on statistical patterns without acquiring rich background knowledge about the physical and social world we live in. Thus far, such systems are not robust when given examples that fall outside the data distribution that they were trained on (Gordon and Van Durme, 2013; Davis and Marcus, 2015; Schubert, 2015).

The Winograd Schema Challenge (WSC), proposed by Levesque et al. (2011) as an alternative to the Turing Test (Turing, 1950), was designed to challenge the dominant paradigm of AI systems that rely on statistical patterns without deep understanding about how the world works. Concretely, Levesque et al. (2011) introduced simple pronoun resolution problems that are trivial for humans but hard for machines by crafting problems not to be easily solvable based on frequent patterns in language. The WSC problems are defined to be a pair (called *twin*) of questions with two answer choices. Here is an example:

1a. Pete envies Martin *because* **he** is successful.
1b. Pete envies Martin *although* **he** is successful.
Question: Is **he** Pete or Martin?
Answers: 1a - Martin, 1b - Pete

These *twin* questions consist of a pair of nearly identical sentences that include *trigger word(s)* that flips the correct answer.

Although WSC questions are carefully crafted by experts, recent studies have shown that they are still prone to incidental biases that statistical methods can exploit. These biases are roughly of two types: (a) *language-based* and (b) *dataset-specific* biases. *Language-based* bias, or *word association* bias (Trichelair et al. (2018)), refers to the case where the correct answer aligns with more frequent patterns in natural language, thus can be easily solved by neural language models trained over large corpora (Table 1 (3) and (4)).

Dataset-specific bias, on the other hand, is the case of *annotation artifacts* or *spurious correlation* that several recent studies have reported on crowdsourced datasets (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018). Importantly, even when an individual instance of a WSC problem is free of language-based bias that the original designers of the WSC intended to avoid, a collection of WSC instances can still contain spurious patterns

		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**

Table 1: WSC examples: (1)-(3) from WSC (Levesque et al., 2011) and (4) from DPR (Rahman and Ng, 2012)). Examples marked with ✗ have language-based bias that today's language models can easily detect.

that can be exploited by statistical models. This type of bias gets introduced as problem authors subconsciously repeat similar problem-crafting strategies, which reveal how particular trigger words, sentence structure, or positive/negative sentiment correlate with the correct answers.

We introduce **WINOGRANDE**,[1] a new collection of WSC problems that are constructed to be robust against both types of biases discussed above. Compared to the original WSC and the variants (§2), WINOGRANDE presents problems that are more challenging by reducing such biases, while also scaling to a significantly larger number of problems (273 to 44k) by crowdsourcing.

Crowdsourcing a large-scale dataset of WSC examples has been considered infeasible primarily due to the pre-requisite knowledge about linguistics and the structural constraints of twin sentences.(Trichelair et al., 2018; Talmor et al., 2018) One novelty of our work is that we demonstrate a method to collect WSC problems at scale through crowdsourcing. We show that the crowdsourced examples maintain the characteristics of WSC; they are easy for humans to answer (above 90% accuracy) but very challenging for state-of-the-art deep neural models. Specifically, we introduce a strategic crowdsourcing design to diversify the context of the problems (§3), followed by introducing a variant of adversarial filtering algorithm (Zellers et al., 2018), AFLITE, for systematically reducing spurious patterns that state-of-the-art statistical approaches can exploit (§4).

While we show that WINOGRANDE is considerably challenging for existing state-of-the-art methods based on pre-trained language models such as BERT Devlin et al. (2018) (§5), we also present that WINOGRANDE provides powerful transfer learning ability to other existing commonsense benchmarks (§6), reporting new state-of-the-art results across several benchmarks, including the original WSC (Levesque et al., 2011) (72.2% → 77.6%), PDP (Morgenstern et al., 2016) (70.0% → 75.0%) DPR (Rahman and Ng, 2012) (76.4% → 86.9%), and COPA (Roemmele et al., 2011) (71.2% → 81.0%). On the Winogender (Rudinger et al., 2018) dataset – which quantifies the gender bias in a trained model – we show that a model trained on WINOGRANDE has significantly lower bias compared to other rule-based and neural models.

Although a substantial increase of the state-of-the-art over multiple challenging benchmarks is exciting, we cautiously argue 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 run the risk of overestimating the true capabilities of machine intelligence on commonsense reasoning. We leave it as a future research question to determine how much of our improvements indicate a true stride in machine commonsense as opposed to a more effective exploitation of biases in datasets.

2 Existing WSC-style datasets

We briefly describe existing WSC-style datasets. Table 2 summarizes them and provides additional statistics about the size, the average token length per sentence and the size of their vocabulary.

WSC (Levesque et al., 2011) This is the original Winograd Schema Challenge dataset, which consisits of 273 problems. The problems are manually crafted by the authors by avoiding word association bias as much as possible (e.g., using the number of search results by Google), although Trichelair et al. (2018) later report that 13.5% of the questions have

[1]The data and codebase are available at https://mosaic.allenai.org/projects/winogrande

Dataset	#Probs	Avg Len	#Vocab
WSC	273	19.1	919
DPR	1,886	15.9	4,127
PDP	80	39.5	594
COPA	1,000	13.3	3,369
Winogender	720	15.6	523
WinoBias	3,168	13.9	1640
SuperGLUE-WSC	804	28.4	1711
WINOGRANDE-debiased	25,680	20.9	14,622
WINOGRANDE-full	43,985	20.7	17,468

Table 2: Existing WSC-style datasets (§2) and relevant statistics (the number of problems, the average sentence length by tokens, and the size of vocabulary). We propose WINOGRANDE described in §3 and 4.

word-association bias.

DPR (Rahman and Ng, 2012) DPR (Definite Pronoun Resolution Dataset) introduces 1,886 additional WSC-style problems authored by 30 undergraduate students. Trichelair et al. (2018) 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.

PDP (Morgenstern et al., 2016) PDP (Pronoun Disambiguation Problems) 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.

In addition to the datasets above, there are some other WSC-style datasets that have slightly different formats but share a similar spirit with WSC.

COPA (Roemmele et al., 2011) This dataset introduces 1,000 problems that share the same motivation as that of WSC in terms of evaluating machine commonsense reasoning, but focus instead on script knowledge. Each problem in this dataset is formulated as a binary choice about cause and effect of given premises, which is not structurally constrained as twins in WSC.

Premise: The man broke his toe.
Question: What was the CAUSE of this?
Hypothesis1: He got a hole in his sock.
Hypothesis2: He dropped a hammer on his foot.

Winogender (Rudinger et al., 2018) This dataset introduces 720 problems focusing on pronouns whose antecedents are either a person referred to by their occupation (e.g., "the doctor") or a secondary participant (e.g., "a patient"). The goal of this dataset is to uncover gender bias in coreference resolution systems.

WinoBias (Zhao et al., 2018) This is a concurrent work with Winogender, aimed at diagnosing gender bias in coreference resolution systems. Although the size is larger than Winogender, WinoBias is evaluated by F-scores (i.e., detecting the span) as a coreference resolution task (Pradhan et al., 2014) instead of the binary choice accuracy as in WSC.

SuperGLUE (Wang et al., 2019) SuperGLUE contains multiple datasets for universal benchmarking across different tasks – one of them is a modified version of WSC. We refer to it as SuperGLUE-WSC to differentiate it from the original WSC. SuperGLUE-WSC aggregates the original WSC, PDP and additional PDP-style examples, and recasts them into True/False binary problems, where a sentence with the target pronoun and an answer candidate is given (e.g., "Pete envies **Martin** because *he* is very successful." Q: Does *he* refer to **Martin**? A: True). Therefore, the number of problems are roughly doubled from WSC and PDP, although the size is still relatively small.

3 Crowdsourcing Twins at Scale

The original WSC problems in Levesque et al. (2011) were carefully crafted by experts in the field of knowledge representation and reasoning, who ensured that the problems were trivial for humans yet hard for AI systems. WSC problems have been considered challenging to crowdsource due to the structural constraints of twins and the requirement of linguistic knowledge – but, contrary to this belief, we present an effective approach to create a large-scale dataset (WINOGRANDE) of WSC problems while maintaining its original properties. Our approach consists of a carefully designed crowdsourcing task followed by a novel adversarial filtering algorithm (§4) that systematically removes biases in the data.

Enhancing Crowd Creativity Creating twin sentences from scratch puts a high cognitive load on crowd workers who subconsciously resort to writing pairs that are lexically and stylistically repetitive. To encourage creativity and reduce cognitive load, we employed *creativity from constraints* (Stokes, 2005) – 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 below), while

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).[2] To prime crowd workers, they were instructed to randomly pick an *anchor* word(s) from a randomly assigned WikiHow article[3] and to ensure that the twin sentences contain the *anchor* word, which remarkably improves diversity of topics in the collected data. Additionally, workers were instructed to keep twin sentence length in between 15 and 30 words while maintaining at least 70% word overlap between a pair of twins.[4] Following the original WSC design, 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 both cases, workers are instructed to avoid language-based bias (word association) as much as possible. In total, we collected 56k questions (i.e., 28k twins).

Data Validation We validated the collected questions, because crowdsourced data often contains noisy results. Each questions is validated by a distinct set of three crowd workers. A question is determined valid if (1) all three workers choose the correct answer option, (2) all three workers agree that the two answer options are not equally plausible and (3) the question cannot be answered just by word association in which local context around the target pronoun is given (e.g., "because **it** was going so fast." (**race car** / school bus)).[5] As a result, 90% of the questions (50k) are deemed valid and we discarded the invalid (6k) questions.

While our crowdsourcing procedure addresses instance-level biases, 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 in datasets.

[2]Our crowdsourcing interface is available at https://mosaic.allenai.org/projects/winogrande.

[3]https://www.wikihow.com/Special:Randomizer

[4]All the workers met minimum qualification in AMT: 99% approval rate, $5k$ approvals, and either US, Canada, UK, Australia, or New Zealand as location. The reward was set to be $0.4 per twin sentences.

[5]For each sentence validation, workers were paid $0.03.

4 Systematic Data Bias Reduction

Bias from annotation artifacts Several recent studies (Gururangan et al., 2018; Poliak et al., 2018; Tsuchiya, 2018) have reported the presence of *annotation artifacts* in large-scale (often crowdsourced) datasets. Annotation artifacts are unintentional patterns in the data that leak information about the target label in an undesired way. Machine learning models can exploit such artifacts to solve instances in a dataset by taking a virtual *shortcut*. While dataset creators can tackle biases that they can identify – e.g. point-wise mutual information (PMI) or conditional probability between a word and an inference class in the Stanford Natural Language Inference (SNLI) corpus (Gururangan et al., 2018; Poliak et al., 2018) – and account for them, these approaches assume that the bias exists in a lexical level. However, it does not deny the existence of other biases derived from structural patterns. Modern machine learning models are endowed with high capacity and also tend to be opaque (often called black boxes), which make identifying the source of bias even more challenging. To tackle these biases that are hard to observe manually, we propose AFLITE – a lightweight algorithmic solution for data bias reduction.

Light-weight adversarial filtering Our approach builds upon the adversarial filtering (AF) algorithm proposed by Zellers et al. (2018), 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 re-training 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 like BERT.

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 BERT (Devlin et al., 2018) 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 50k instances in total) to fine-tune BERT (referred to as $\text{BERT}_{\text{embed}}$). We use $\text{BERT}_{\text{embed}}$ to pre-compute

Algorithm 1: AFLITE

Input: dataset $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, ensemble size n, training set size m, cutoff size k, filtering threshold τ
Output: dataset $\mathcal{D}'$

$\mathcal{D}' = \mathcal{D}$
while $|\mathcal{D}'| > m$ **do**
  `// Filtering phase`
  forall $e \in \mathcal{D}'$ **do**
    Initialize the ensemble predictions $E(e) = \emptyset$
  for *iteration* $i : 1..n$ **do**
    Random partition $(\mathcal{T}_i, \mathcal{V}_i)$ of $\mathcal{D}'$ s.t. $|\mathcal{T}_i| = m$
    Train a linear classifier $\mathcal{L}$ on $\mathcal{T}_i$
    forall $e = (\mathbf{x}, y) \in \mathcal{V}_i$ **do**
      Add $\mathcal{L}(\mathbf{x})$ to $E(e)$
  forall $e = (\mathbf{x}, y) \in \mathcal{D}'$ **do**
    $score(e) = \frac{|\{p \in E(e) \text{ s.t. } p=y\}|}{|E(e)|}$
  Select the top-k elements $\mathcal{S}$ in $\mathcal{D}'$ s.t. $score(e) \geq \tau$
  $\mathcal{D}' = \mathcal{D}' \setminus \mathcal{S}$
  if $|\mathcal{S}| < k$ **then**
    break
return $\mathcal{D}'$

the embeddings for the rest of the instances (44k) 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 $\text{BERT}_{\text{embed}}$ is strongly indicative of the correct answer option. If so, we discard the corresponding instances and proceed iteratively.

Algorithm 1 provides the implementation of AFLITE. The algorithm takes as input the *pre-computed* embeddings $\mathbf{X}$ and labels $\mathbf{y}$, along with the size n of the ensemble, the training size m for the classifiers in the ensemble, the size k 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 prediction 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 $m = 15,000$, $n = 64$, $k = 500$, and $\tau = 0.75$.

This approach is also reminiscent of recent work in NLP on adversarial learning (Chen and Cardie, 2018; Belinkov and Bisk, 2018; Elazar and Goldberg, 2018). Belinkov et al. (2019) propose 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, while the hypothesis-only bias is an insightful bias in NLI, we make no assumption about the possible sources of bias in WINOGRANDE. Instead, we adopt a more proactive form of bias reduction by relying on 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 PMI-based filtering. In random data reduction, we randomly subsample the dataset to evaluate how a decrease in dataset size affects the bias. In PMI-based filtering, we first compute the difference (f) of PMIs for each twin (t) as follows:

$$f_t(t_1, t_2) = \sum_{w \in t_1} \text{PMI}(y; w) - \sum_{w \in t_2} \text{PMI}(y; w).$$

Then, we select twins in increasing order of f_t, assuming that higher values of f_t lead to less challenging twin instances.[6]

Figure 1 plots BERT pre-computed embeddings whose dimension is reduced to 2D (*top*) and 1D (*bottom*) using Principal Component Analysis (PCA). We observe that $\text{WINOGRANDE}_{\textit{all}}$ and the two baselines exhibit distinct components between the two correct answer options (i.e., $y \in 1, 2$), whereas such distinction disappears in $\text{WINOGRANDE}_{\textit{debiased}}$, 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 ($0.66 \rightarrow 0.65$). It is interesting to see that PMI-filtering marginally reduces the KL divergence ($0.66 \rightarrow 0.46$), although the principal component analysis on the PMI-filtered subset still leads to a significant separation between the labels. On the other hand, in $\text{WINOGRANDE}_{\textit{debiased}}$, AFLITE reduces the KL divergence dramatically ($0.66 \rightarrow 0.02$) which suggests that this debiased dataset should be challenging for statistical models that solely rely on spurious correlation.

[6] We also evaluated other variations of PMI-filtering such as the absolute difference ($|f|$), maximum PMI ($= \max(\max_{w \in t_1} \text{PMI}(y; w), \max_{w \in t_2} \text{PMI}(y; w))$), and second-order $\text{PMI}(y; w_1, w_2 \in t)$, but we did not observe a significant difference.

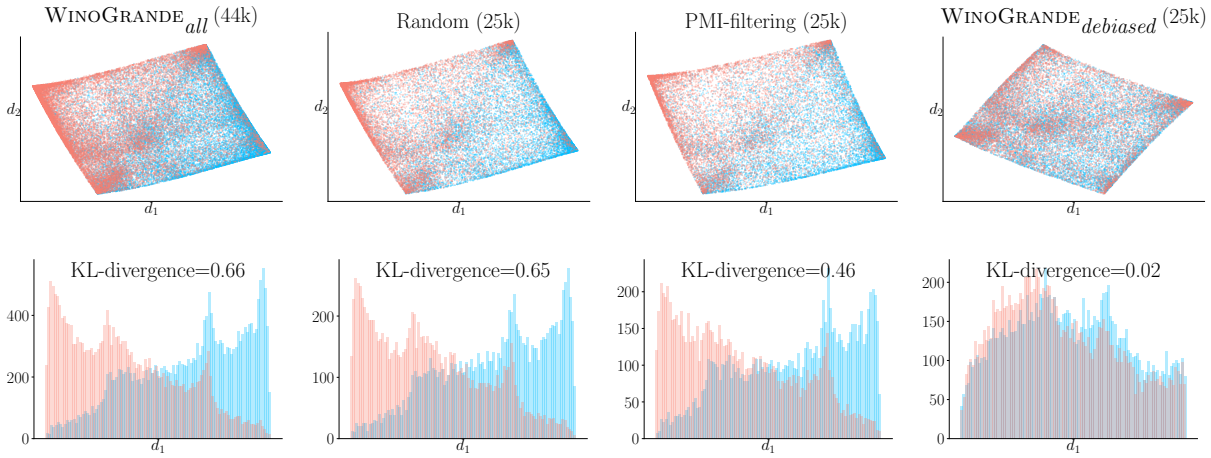

Figure 1: The effect of debiasing by AFLITE. BERT pre-computed embeddings (applied PCA for dimension reduction) are shown in 2D-histograms (*top row*) and 1D-histograms (*bottom row*) for WINOGRANDE$_{all}$, the random samples, PMI-filtered subset, and AFLITE-filtered subset. Data points are colored depending on the label (i.e., the answer y is option 1 (blue) or 2 (red)). In the 1D representation, we show the KL-divergence between $p(d_1, y{=}1)$ and $q(d_1, y{=}2)$.

	Twin sentences	Options (**answer**)
✗	The rock kept its balance on the mountain but the log tumbled down, because **it** was *better* situated for stability.	**rock** / log
	The rock kept its balance on the mountain but the log tumbled down, because **it** was *poorly* situated for stability.	rock / **log**
✗	Nick did not enjoy watching golf as much as Randy because **he** *never* played the game.	**Nick** / Randy
	Nick did not enjoy watching golf as much as Randy because **he** *often* played the game.	Nick / **Randy**
✓	The pizza was warmer than the hot dog because **it** was in the oven for a *longer* amount of time.	**pizza** / hot dog
	The pizza was warmer than the hot dog because **it** was in the oven for a *shorter* amount of time.	pizza / **hot dog**
✓	Sarah accused Katrina of cheating by looking at her cards, because **she** kept *losing* the game.	**Sarah** / Katrina
	Sarah accused Katrina of cheating by looking at her cards, because **she** kept *winning* the game.	Sarah / **Katrina**

Table 3: Examples that have *dataset-specific* bias detected by AFLITE (marked with ✗). The words that include (dataset-specific) polarity bias (§4) are highlighted (positive and negative). For comparison, we show examples selected from WINOGRANDE$_{debiased}$ (marked with ✓).

What bias has been actually detected by AFLITE? Is the bias really spurious and undesirable according to the original WSC's goal? Table 3 presents examples of structural biases (i.e., spurious relation) that AFLITE has detected as a dataset-specific bias. We see a structural pattern in the first two twins, where the local context (or sentiment) between the answer option and the target pronoun are highly correlated. In other words, these problems can be easily answered by simply looking at the surrounding context and the polarity of the sentiment (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 (Gururangan et al., 2018; Poliak et al., 2018), and it is hard to detect these biases using heuristics such as PMI-filtering. Instead of depending on such heuristics, AFLITE is able to detect samples that potentially have such biases algorithmically.

After applying the AFLITE algorithm, we obtain a *debiased* dataset of $25,680$ instances split into training ($18,538$), development ($2,863$), and test ($4,279$) sets.

5 Experimental Analysis

5.1 Benchmark Models

We evaluate the WINOGRANDE$_{debiased}$ on methods/models that have been effective on the original WSC.

Wino Knowledge Hunting Wino Knowledge Hunting (WKH) by Emami et al. (2018) 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. This IR-oriented

approach comes from an important line of work in coreference resolution (Kobdani et al., 2011; Ratinov and Roth, 2012; Bansal and Klein, 2012; Zheng et al., 2013; Peng et al., 2015; Sharma et al., 2015).

Ensemble Neural LMs Trinh and Le (2018) is one of the first attempts to apply a neural language model which is pre-trained on a very large corpora (including 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.

OpenAI-GPT OpenAI-GPT (Radford et al., 2018) is one of the earliest methods that uses large-scale pre-trained neural language modeling. While the first version of OpenAI-GPT did not report its performance on WSC,[7] the updated model (Radford et al., 2019) reports 70.7% on the original WSC.

BERT BERT (Devlin et al., 2018) is another pre-trained neural model which has bidirectional paths and consecutive sentence representations in hidden layers. We use three different BERT-related models: 1) BERT masked-LM (BERT-lm), 2) BERT-single-finetuning (BERT-ft), and 3) BERT-sequential-finetuning (BERT-seqft). For BERT-lm, we use the pre-trained BERT-large model as a language model by comparing the likelihood of each candidate answer. For BERT-ft, we split the sentence into context and option using the candidate answer as delimiter. The input format becomes `[CLS] context [SEP] option [SEP]`; e.g., *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). For BERT-seqft, we first finetune BERT-large on an auxiliary dataset (DPR in our case), and then fine-tune the resulting pre-trained model on the target dataset. We used grid-search for hyper-parameter tuning: learning rate $\{1e-5, 3e-5, 10e-5\}$, number of epochs $\{3, 4, 5, 10\}$, batch-size $\{4, 8, 16\}$.

Word association baseline Using BERT-seqft, 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 ($w_{t-2:\text{EOS}}$) surrounding the blank to be filled (w_t). This is analogous to the *hypothesis-only* baseline in NLI (Poliak et al., 2018), where the task (dataset) does not require the full context to achieve high performance. In order to see if there is an effect of AFLITE for language-based bias, we fine-tune BERT with either randomly selected samples (25k) or debiased samples (25k), subsequently referred to as $\text{seqft}_{\text{random}}$ and $\text{seqft}_{\text{debiased}}$, respectively.

Human evaluation In addition to the methods described above, we compute human performance as the majority vote of three crowd workers for each question. We find that AFLITE does not adversely effect the quality of the dataset as humans are still able to achieve over 90% accuracy, significantly higher than the performance of the best model (62%).

Methods	dev acc. (%)	test acc.(%)
Random	50.0	50.0
WKH	50.8	50.1
Ensemble LMs	53.0	50.9
OpenAI-GPT	61.6	50.7
BERT-lm	53.5	51.8
BERT-ft	53.0	52.3
BERT-seqft	**63.5**	**61.6**
local context ($\text{seqft}_{\text{random}}$)	56.1	54.1
local context ($\text{seqft}_{\text{debiased}}$)	57.0	55.4
Human Perf.	92.0	90.8

Table 4: Performance of several baseline systems on WINOGRANDE$_{debiased}$. The best performing model (BERT-seqft) is over 28 percentage points below human performance.

5.2 Results

Table 4 shows the results. Most baselines only achieve chance-performance, while the best model, BERT-seqft achieves 61.6% test-set accuracy. Crowd workers achieve 90.8% test-set accuracy, indicating that the WINOGRANDE$_{debiased}$ is still easy for humans to answer as desired. The large gap between the performance of the best model and that of humans provides significant scope of improvements for future research. Additionally, word association (i.e., local context) baselines ($\text{seqft}_{\text{random}}$ and $\text{seqft}_{\text{debiased}}$) achieve close to chance-level performance, illustrating that WINOGRANDE$_{debiased}$ cannot be answered by the local context only. It is interesting to see that there is no performance gap between $\text{seqft}_{\text{random}}$ and

[7] Instead, the model was evaluated on the WNLI dataset from the GLUE benchmark, although it did not perform as well as the baseline.

WSC (Levesque et al., 2011)	
Liu et al. (2016)	52.8
WKH (Emami et al., 2018)	57.1
Ensemble LMs (Trinh and Le, 2018)	63.8
GPT2 (Radford et al., 2019)	70.7
BERT-ft (Kocijan et al., 2019)	72.2
This work	**77.6**
Humans (Bender, 2015)	92.1
Humans*	96.5
PDP (Morgenstern et al., 2016)	
Liu et al. (2016)	61.7
Trinh and Le (2018)	70.0
This work	**75.0**
Humans (Davis et al., 2016)	90.9
Humans*	92.5
DPR (Rahman and Ng, 2012)	
Rahman and Ng (2012)	73.0
Peng et al. (2015)	76.4
This work	**86.9**
Humans*	95.2
COPA (Roemmele et al., 2011)	
Gordon et al. (2011)	65.4
Sasaki et al. (2017)	76.4
This work	**81.0**
Humans (Gordon et al., 2012)	99.0

Table 5: Accuracy (%) on existing WSC-related tasks. We ran human evaluation with our crowd worker pool (indicated by *).

Winogender (Rudinger et al., 2018)					
	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		
BERT_{WSC}	No	44.2	64.2	17.5	19.2
	Yes	61.7	45.0		
BERT_{DPR}	No	58.3	44.2	13.3	16.6
	Yes	45.0	60.8		
$\text{BERT}_{\text{WG-deb}}$	No	69.2	65.0	4.2	3.3
	Yes	65.0	68.3		
$\text{BERT}_{\text{WG-all}}$	No	83.3	76.7	6.6	4.1
	Yes	76.7	80.8		

Table 6: Accuracy (%) and gender bias on Winogender dataset. "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. (2018); RULE is Lee et al. (2011), STATS is Durrett and Klein (2013), and NEURAL is Clark and Manning (2016). BERT_{X} corresponds to the BERT-large model fine-tuned on X, where X is either WSC, DPR, $\text{WINOGRANDE}_{debiased}$, or WINOGRANDE_{all}.

$\text{seqft}_{\text{debiased}}$. This indicates that the word association bias has already been removed during the data validation process (§3).

6 Using WINOGRANDE as a Resource

WINOGRANDE contains a large number of WSC style questions. In addition to serving as a benchmark dataset, we use WINOGRANDE_{all} 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, DPR, COPA, and Winogender). We establish state-of-the-art results across several of these existing benchmark datasets.

Experimental Setup Our model is based on BERT finetuned with WINOGRANDE_{all} and the hyper-parameters are determined by the following. For WSC, we used PDP as the dev set to choose the best hyper-parameter set, and vice versa (i.e., WSC as the dev set for PDP). Since DPR and COPA provide training set, we used it as a dev set to determine the hyper parameter set to evaluate the test set. For hyper parameter search, we use the same grid search strategy as in §5. Winogender dataset provides a test set only, and we use the WINOGRANDE_{all} dev set as a proxy.

Additional Human Evaluation We also report human performance for WSC, PDP, and DPR to check the quality of our crowd worker pool as well as supporting previous findings. To our knowledge, this is the first work to report human performance on DPR dataset.[8]

Results The results are shown in Table 5 and Table 6. Overall, BERT finetuned with WINOGRANDE_{all} helps improve the accuracy of all the related tasks (Table 5). At first glance, these improvements may not seem surprising because WINOGRANDE_{all} can be regarded as additional training data for each dataset (particularly WSC, PDP, and DPR). However, the improvement on the COPA

[8] We didn't run human evaluation on COPA and Winogender because they have slightly different question formats from WSC, PDP, DPR, and WINOGRANDE.

dataset (76.4% → 81.0%) is not explained by the same logic, because the COPA task is not a pronoun resolution task like the Winograd Schema Challenge. This indicates that our WINOGRANDE$_{all}$ can serve as a resource to support commonsense knowledge transfer.

Important Implications We consider that while 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.

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.

If ΔF or ΔM is large, it indicates that the model is highly gender-biased, whereas $|\Delta F| = |\Delta M| = 0$ (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 6. We find that the BERT model trained on WINOGRANDE (BERT$_{\text{WG-debiased}}$, BERT$_{\text{WG-full}}$) both demonstrate considerably smaller gender-bias ($|\Delta F|$ and $|\Delta M|$) compared to the BERT trained on other datasets. It is important to note that the difference comes purely from dataset but not the model structure with pretraining. Does the data size correlate with the reduction of gender gap? This may be true but is not always the case. The gender gap in BERT$_{\text{WG-debiased}}$ (25k) is smaller than that in BERT$_{\text{WG-full}}$ (44k), which indicates a possibility that AFLITE can reduce undesirable gender bias in the dataset in addition to reducing structural biases (§4).

7 Conclusions

We introduce WINOGRANDE, a new collection of WSC problems that is significantly larger than existing variants of the WSC dataset. WINOGRANDE consists of 44k instances, half of which determined adversarial. To create a dataset that is robust against spurious statistical biases, we also present AFLITE – a novel light-weight adversarial filtering algorithm. The resulting dataset is significantly more challenging for existing state-of-the-art models while being trivially easy for humans.

Using WINOGRANDE as a resource, we demonstrate effective transfer learning and achieve state-of-the-art results on several WSC-style benchmark datasets. While this is an exciting result, we also discuss the risk of overestimating the performance of the existing state-of-the-art methods on the existing commonsense benchmarks. There is a possibility that they contain spurious statistical patterns (annotation artifacts) that leak information about the target label in an undesirable way.

We advocate for 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 the next decade or two, we propose that future benchmarks should *dynamically evolves together with the evolving state-of-the-art.*

References

Mohit Bansal and Dan Klein. 2012. Coreference semantics from web features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 389–398, Jeju Island, Korea. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. *ICLR*, abs/1711.02173.

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In *Joint Conference on Lexical and Computational Semantics (StarSem)*.

David Bender. 2015. Establishing a human baseline for the winograd schema challenge. In *Proceedings of the 26th Modern Artificial Intelligence and Cognitive Science Conference (MAICS)*, pages 39–45.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *NAACL-HLT*.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Ernest Davis, Leora Morgenstern, and Charles Ortiz. 2016. Human tests of materials for the winograd schema challenge 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *EMNLP*.

Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. A generalized knowledge hunting framework for the winograd schema challenge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 25–31, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Andrew S. Gordon, Cosmin Adrian Bejan, and Kenji Sagae. 2011. Commonsense causal reasoning using millions of personal stories. In *AAAI*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

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

Hamidreza Kobdani, Hinrich Schuetze, Michael Schiehlen, and Hans Kamp. 2011. Bootstrapping coreference resolution using word associations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 783–792, Portland, Oregon, USA. 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, ACL 2019, Florence, Italy, July 28 - August 2, 2019*. Association for Computational Linguistics.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA. Association for Computational Linguistics.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, page 47.

Quan Liu, Hui Jiang, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2016. Commonsense knowledge enhanced embeddings for solving pronoun disambiguation problems in winograd schema challenge. *arXiv preprint arXiv:1611.04146*.

Leora Morgenstern, Ernest Davis, and Charles L Ortiz. 2016. Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, 37(1):50–54.

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, Denver, Colorado. Association for Computational Linguistics.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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

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

Lev Ratinov and Dan Roth. 2012. Learning-based multi-sieve co-reference resolution with knowledge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1234–1244, Jeju Island, Korea. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Shota Sasaki, Sho Takase, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2017. Handling multiword expressions in causality estimation. In *IWCS*.

Lenhart K Schubert. 2015. What kinds of knowledge are needed for genuine understanding. In *IJCAI 2015 Workshop on Cognitive Knowledge Acquisition and Applications (Cognitum 2015)*.

Arpit Sharma, Nguyen H. 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 *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1319–1325. AAAI Press.

Patricia D Stokes. 2005. *Creativity from constraints: The psychology of breakthrough*. Springer Publishing Company.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

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

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

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Alan M Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–433.

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

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jiaping Zheng, Luke Vilnis, Sameer Singh, Jinho D. Choi, and Andrew McCallum. 2013. Dynamic knowledge-base alignment for coreference resolution. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 153–162, Sofia, Bulgaria. Association for Computational Linguistics.