A Dataset construction

A.1 Switching candidates

This dataset contains the original WSC with the switched version of each sentence whenever the process does not obscure the sentence or affect the rationale used to resolve the target pronoun. To construct this dataset, we first automatically switch the two candidates.

(1) **Original sentence** *Emma* did not pass the ball to *Janie* although she saw that she was open.

(2) **Switched sentence** *Janie* did not pass the ball to *Emma* although she saw that she was open.

This process can make a sentence obscure, as in the following example:

(3) **Original sentence** Sam broke both his *ankles* and he's walking with *crutches*. But a month or so from now they should be better.

(4) **Switched sentence** Sam broke both his *crutches* and he's walking with *ankles*. But a month or so from now they should be better.

The sentence obtained is not correct as *walking with ankles* is neither semantically correct nor requires the same resolution rationale. To filter out these sentences, we asked three English native speakers, who did not have prior knowledge on the WSC, to classify the sentences as *Switchable* or *Not Switchable*. We keep the switched version of the sentence if the three annotators agreed. This procedure produces a dataset of 131 switched sentences with a high agreement as shown in Table 1.

A.2 *Associativity*

This dataset contains the original WSC sentences labeled as *associative* or *non-associative*. Associative Winograd sentences are those in which one candidate antecedent associates strongly with the clause containing the pronoun, while the other candidate antecedent exhibits no such association strength. For example:

(5) In the storm, *the tree* fell down and crashed through *the roof* of my house. Now, I have to get [it] repaired.

Here, *the roof* can be argued to be much more strongly associated with *repaired*, and on this basis, can be used to resolve the pronoun.

An example of a non-associative sentence is:

(6) Everyone really loved *the oatmeal cookies*; only a few people liked *the chocolate chip cookies*. Next time, we should make more of [them] .

Here, we don't expect, at least *a priori*, that *oatmeal cookies* associate more than *the chocolate chip cookies* with the clause, *"we should make more of them"* and therefore can be argued to be much more robust to techniques that rely on cooccurence statistics.

We split the WSC into smaller associative and non-associatve datasets by conducting a human study similar to that in A.1. The three annotators only had access to the clause containing the pronoun (e.g. *get [it] repaired* and *Next time, we should make more of [them]* for (5) and (6) respectively), and the two candidate antecedents. Using these, they were asked to categorize a sentence as associative or non-associative according to whether or not they saw a strong association between one entity and the clause, and no such association with the other entity. We chose to consider a sentence as *associative* if the three annotators unanimously agreed. This process lead to a high inter-annotator agreement as shown in Table 1 and resulted in an *associative* dataset with 37 sentences and a *non-associative* dataset with 252 sentences (there were 42 sentences for which there was not a full agreement).

B Lucky draw

We consider a random classifier so that for each sentence, it chooses one of the two candidates. Since the dataset is balanced, the probability of getting the correct answer is 50%. When classifying the 273 instances, the number of correct answers X is a binomial random variable. The probability of getting more than 55% accuracy (more than 150 correct answers) is given by:

Statistic used	Score Switchability	Score Associativity
Fleiss' Kappa	0.96	0.79

Table 1: Inter-rater agreement measured using Fleiss's Kappa for both the switching and the associativity annotations

$$P(X > 150) = 1 - P(X \leq 150)$$
$$P(X > 150) = 1 - \sum_{i=0}^{150} P(X = i)$$
$$P(X > 150) = 1 - \sum_{i=0}^{150} \binom{273}{i} 0.5^i (1 - 0.5)^{273-i}$$
$$P(X > 150) = 1 - 0.5^{273} \sum_{i=0}^{150} \binom{273}{i}$$
$$P(X > 150) = 0.04$$

It shows that the probability of scoring more than 55% on the WSC using a random classifier is 4%. When repeating the experiments 10 times, the probability that one of the experiments gives an accuracy greater than 55% corresponds to $1 - P(X \leq 150)^{10} = 0.37$. Practically, on the WSC, this means that if we have a pool of 10 random classifiers, there is more than a 1-in-3 chance that one of them scores more than 55%.

C Implementation Details

For the WSC, we reproduced the results for the language model and the Knowledge-Hunter using the authors' released code available on Github;
The language model:
https://github.com/tensorflow/models/tree/master\/research/lm_commonsense
The Knowledge Hunter:
https://github.com/aemami1/Wino-Knowledge-Hunter.
For GPT-2, we use the implementation released in the paper and slighly modified it. We have attached the implementation with the submission.
For BERT, we have attached the implementation with the submission. The modifications we have made to the original implementation include the necessary adapations for SWAG.