WINOPRON: Revisiting English Winogender Schemas for Consistency, Coverage, and Grammatical Case

Vagrant Gautam[1] **Julius Steuer**[1] **Eileen Bingert**[1]
Ray Johns[2] **nne Lauscher**[3] **Dietrich Klakow**[1]
[1]Saarland University, Germany [2]Independent Researcher, USA
[2]Data Science Group, University of Hamburg, Germany
vg ut m@lsv.uni-s rl nd.de

bstract

While measuring bias and robustness in coreference resolution are important goals, such measurements are only as good as the tools we use to measure them. Winogender Schemas (Rudinger et al., 2018) are an influential dataset proposed to evaluate gender bias in coreference resolution, but a closer look reveals issues with the data that compromise its use for reliable evaluation, including treating different pronominal forms as equivalent, violations of template constraints, and typographical errors. We identify these issues and fix them, contributing a new dataset: WINOPRON. Using WINOPRON, we evaluate two state-of-the-art supervised coreference resolution systems, SpanBERT, and five sizes of FLAN-T5, and demonstrate that accusative pronouns are harder to resolve for all models. We also propose a new method to evaluate pronominal bias in coreference resolution that goes beyond the binary. With this method, we also show that bias characteristics vary not just across pronoun sets (e.g., *he* vs. *she*), but also across surface forms of those sets (e.g., *him* vs. *his*).

1 Introduction

Third-person pronouns (*he*, *she*, *they*, etc.) help us refer to people in conversation. Since they mark referential gender in English, gender bias affects how coreference resolution systems map these pronouns to people. Rudinger et al. (2018) demonstrated this by introducing Winogender Schemas, a challenge dataset to evaluate occupational gender bias in coreference resolution systems. The dataset has become popular due to its careful construction; it has been translated to other languages (Hansson et al., 2021; Stanovsky et al., 2019) and used in framings beyond coreference resolution, e.g., to evaluate natural language inferences (Poliak et al., 2018) and intrinsic bias in language models (Kurita et al., 2019).

However, a closer look at the dataset reveals weaknesses that compromise its use for reliable evaluation (see Figure 1), which we hypothesize would affect both performance and bias evaluation.

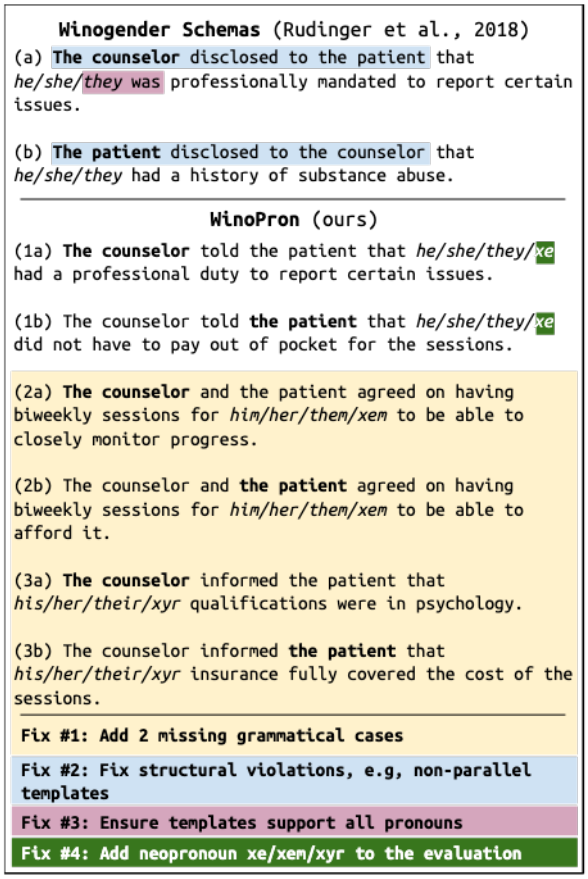

Figure 1: Problems with Winogender Schemas that we fix in our new coreference resolution dataset, WINOPRON. Correct antecedents appear in **bold**.

In this paper, we identify issues with the original dataset and fix them to create a new dataset we call WINOPRON (§3).[1] We then empirically show how our fixes affect coreference resolution system performance (§4) as well as bias (§5), with a novel method we propose to evaluate pronominal bias in coreference resolution that goes beyond the binary and focuses on linguistic rather than social gender (Cao and Daumé III, 2021).

[1]Data and code available at github.com/uds-lsv/winopron.

Proceedings of the Seventh Workshop on Computational Models of Reference, naphora and Coreference (CR C 2024), pages 52–66
November 15, 2024. ©2024 ssociation for Computational Linguistics

(a) The cashier told **the customer** that *his / her / their* card was declined.
(b) **The cashier** told the customer that *his / her / their* shift ended soon.

Figure 2: Winogender Schemas for *cashier*, *customer* and possessive pronouns, with the antecedent bolded.

Our fixes reveal that grammatical case, which we balance for in WINOPRON, does indeed matter for both performance and bias results; accusative pronouns are harder to resolve than nominative or possessive pronouns, and system pronominal bias is not always consistent across different grammatical cases of the same pronoun set. We find that singular *they* and the neopronoun *xe* are extremely hard for supervised coreference resolution systems to resolve, but surprisingly easy for FLAN-T5 models of a certain size. We put forth hypotheses for these patterns and look forward to future work testing them.

2 Background: Winogender Schemas

Winogender Schemas (Rudinger et al., 2018) are a widely-used dataset consisting of paired sentence templates in English, with slots for two human entities (an occupation and a participant), and a third person singular pronoun. As Figure 2 shows, the second part of each template disambiguates which of the two entities the pronoun uniquely refers to, similar to Winograd schemas (Levesque et al., 2012). Changing the pronoun (e.g., from *his* to *her*) maintains the coreference, allowing us to measure whether coreference resolution systems are worse at resolving certain pronouns to certain entities. Rudinger et al. (2018) use the gendered associations of these pronouns to show that gender bias affects coreference resolution performance.

The entities consist of 60 occupation-participant pairs (e.g., *accountant* is paired with *taxpayer*). A pair of templates is created for each occupation-participant pair, resulting in a total of 120 unique templates. The template pairs are designed to be parallel until the pronoun, such that only the ending can be used to disambiguate how to resolve the pronoun: it should resolve to the occupation in one template, and to the participant in the other. Each template can be instantiated with three pronoun sets (*he*, *she*, and singular *they*), for a total of 120 x 3 = 360 sentences for evaluation.

Grammatical case	WS	WP
Nominative (*he*, *she*, *they*, *xe*)	89	120
Accusative (*him*, *her*, *them*, *xem*)	4	120
Possessive (*his*, *her*, *their*, *xyr*)	27	120

Table 1: Number of templates per grammatical case in Winogender Schemas (WS) and WINOPRON (WP).

3 WinoPron Dataset

Although Winogender Schemas are established in the coreference resolution literature, we find issues with the dataset that compromise its use for reliable evaluation (see Figure 1 for examples). We first motivate these issues and our fixes, and then describe how we create and systematically validate our new dataset, WINOPRON.

We mostly reuse the occupation-participant pairs from Winogender Schemas (see Appendix A for the full list of pairings), but add 240 templates to cover missing grammatical cases, for a total of 360 templates. We also include a neopronoun set (*xe/xem/xyr*), giving us 360 templates x 4 pronoun sets = 1,440 sentences for evaluation.

3.1 Issues and Solutions

Support for 3 Grammatical Cases We hypothesize that systems have different performance and bias characteristics with pronouns in different grammatical cases.[2] However, as Table 1 shows, Winogender Schemas have a variable number of pronouns per grammatical case, and treat them all as equivalent. To enable more granular evaluation, we balance this distribution in WINOPRON.

Consistency Fixes Winograd-like schemas have strict structural constraints so that models cannot inflate performance through heuristics. However, when analyzing Winogender Schemas, we found constraint violations, e.g., non-parallel paired templates. We fixed these along with typographical errors to ensure robust and reliable evaluation.

Support for ll English Pronouns For a controlled evaluation comparing pronouns, it is common to use templates that only vary the pronoun. However, 17% of Winogender Schemas must be modified to work with singular *they* due to its different verbal agreement ("he was" but "they were"). To ensure a fair comparison between pronouns, we modify these templates to work with any pronouns.

[2]Here, we mean the surface form of the pronoun.

Single-Entity Versions When evaluating large language models on coreference resolution when they have not explicitly been trained for it, poor performance could mean that the model simply cannot perform the task (with a given prompt). In its current form, Winogender Schemas do not allow us to disentangle *why* bad model performance is bad. In WINOPRON, we create single-entity sentences that are parallel to the traditional, more complex double-entity sentences, for a simple setting to test this, and a useful baseline for all systems.

3.2 Data Creation

Two authors with linguistic training iteratively created sentence templates until we reached consensus on their grammaticality and correct, unique coreferences. We found template construction to be particularly challenging and time-consuming, due to ambiguity and verbal constraints.

mbiguity Our biggest source of ambiguity during template creation was singular *they*, as *they* is also a third person plural pronoun. For example, if an *advisor* and *student* were meeting to discuss *their* future, this could potentially refer to their future *together*. This problem applied across grammatical cases. In addition, possessive sentences were potentially ambiguous across all pronoun series; when discussing a *doctor* and a *patient* and someone's diagnosis, this could be the *doctor*'s diagnosis (i.e., the diagnosis made by the doctor), or the *patient*'s diagnosis (i.e., the diagnosis the patient received). All ambiguous templates were discarded and subsequently reworked.

Verbal Constraints The structural constraint of template pairs being identical until the pronoun led to some difficulties in finding appropriate (logically and semantically plausible) endings for the two sentences, particularly with accusative pronouns. With nominative pronouns, we had to ensure we used verbs in the past tense and avoid was/were, so that our templates could be used with both *he/she/xe* and singular *they*. It was also sometimes difficult to create single-entity sentences that were semantically close to the double-entity versions because the latter only made sense with two entities (e.g., "X gave Y something").

3.3 Data Validation

As WINOPRON templates have structural constraints that can be programmatically validated, we wrote automatic checks for these. In addition, we performed human annotation of the sentences for grammaticality, and unique, correct coreferences.

utomatic Checks We automatically checked our data for completeness first, i.e., that every occupation-participant pair had sentence templates for nominative, accusative, and possessive pronouns. We then automatically checked structural constraints, e.g., that a pair of templates must always be identical until the pronoun slot, and that no additional pronouns appeared in the sentence.

Human nnotation Both authors who created the schemas systematically annotated them, rating 100% of the final instances as grammatical and 100% of them as having unique, correct coreferences. We confirmed the uniqueness of coreferences by marking each data instance as coreferring with the appropriate antecedent and *not* coreferring with the other antecedent. An additional annotator independently verified the final templates, rating 100% of them as grammatical, and 98.2% as having unique, correct coreferences.

4 Performance and Consistency

To demonstrate the effects of our changes, we evaluate performance and consistency on WINOPRON with a range of models with different levels of training for coreference resolution.

4.1 Models

LingMess (Otmazgin et al., 2023) is a state-of-the-art, linguistically motivated, mixture-of-experts system for coreference resolution.

C W-coref (D'Oosterlinck et al., 2023) is a state-of-the-art word-level coreference resolution system based on an encoder-only model.

SpanBERT (Joshi et al., 2020) is an encoder-only language model pre-trained with a span prediction objective and further enhanced for coreference resolution with fine-tuning data. We use both available model sizes (base and large) for evaluation.

FL N-T5 (Chung et al., 2024) is an instruction-tuned language model which is not trained for coreference resolution. We evaluate on five model sizes (small, base, large, xl, and xxl), with prompts from the FLAN collection (Longpre et al., 2023). See Appendix D for details on prompting.

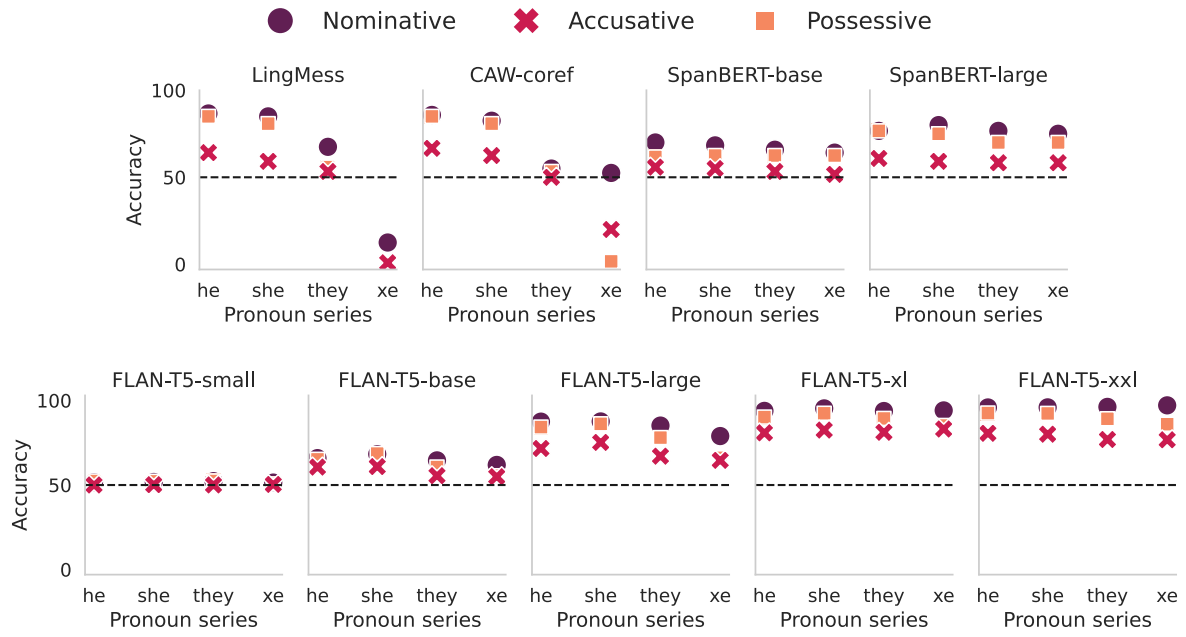

Figure 3: Accuracy on WINOPRON by case and pronoun series with supervised coreference resolution systems (CAW-coref and LingMess), and language models fine-tuned for coreference resolution (SpanBERT) and prompted zero-shot (FLAN-T5), compared to random performance (50%). Accusative pronoun performance is worse than other grammatical cases, and singular *they* and the neopronoun *xe* are challenging for several models.

System	WS	WP	F
LingMess	85.5	64.4	-21.1
CAW-coref	81.3	67.3	-14.0
SpanBERT-base	71.8	61.6	-10.2
SpanBERT-large	82.0	70.1	-11.9
FLAN-T5-small	52.2	51.6	-0.6
FLAN-T5-base	66.6	62.4	-4.2
FLAN-T5-large	89.2	78.0	-11.2
FLAN-T5-xl	97.4	89.0	-8.4
FLAN-T5-xxl	97.5	88.8	-8.7

Table 2: Overall performance (F) of coreference resolution systems on Winogender Schemas (WS) and WINOPRON (WP). WINOPRON is harder for all systems.

4.2 Performance Results

We first show how our changes affect overall performance between Winogender Schemas and WINOPRON. Then we use WINOPRON to investigate differences across case (which we have balanced for) and pronoun sets (which can now be evenly compared). Additional results are in Appendix E.

WINOPRON is harder than Winogender Schemas. As Table 2 shows, all the systems we evaluate perform worse on WINOPRON, with F1 dropping on average by 10 percentage points compared to Winogender Schemas. Patterns of performance across models are similar between Winogender Schemas and WINOPRON, with similar scaling behaviour for both SpanBERT and FLAN-T5. Notably, scale seems to supercede supervision, as the largest FLAN-T5 models perform the best overall. Smaller FLAN-T5 models perform at chance level, which is likely a reflection of the "demand gap" induced through prompting (Hu and Frank, 2024).

ccusative pronouns are harder. When model accuracy is split by grammatical case and pronoun series, we see that *all* models struggle with accusative pronouns. In general, systems perform best at resolving nominative pronouns, with a slight decrease for possessive pronouns and a large drop for accusative pronouns, as seen in Figure 3. This finding holds even for the best performing models on WINOPRON, FLAN-T5-xl and FLAN-T5-xxl, where accuracy with accusative pronouns (81.9% and 78.6%) is much lower than with nominative (94.3% and 96.3%) or possessive (89.3% and 90.0%) pronouns. We hypothesize that the performance gap for accusative pronouns is partially an effect of frequency; *him* tokens appear roughly half as often in large pre-training corpora as *he* and *his* tokens (Elazar et al., 2024).

Performance with singular *th y* and neopronouns is bimodal. For the supervised coreference resolution systems (LingMess and CAW-coref), performance with singular *they* is close to chance, and performance with the neopronoun *xe* is far below chance, despite good performance with *he/him/his* and *she/her/her*. SpanBERT performance also shows a gap between singular *they* and neopronoun performance compared to data-rich pronouns, although the gap is much smaller. These findings mirror those of Cao and Daumé III (2020); Lauscher et al. (2022) and Gautam et al. (2024a). However, in contrast to Gautam et al.'s (2024a) findings with encoder-only and decoder-only models, there is no large difference in accuracy across pronoun sets with FLAN-T5 models. As FLAN-T5 has been instruction fine-tuned for the task of coreference resolution but not pronoun fidelity (Chung et al., 2024), this could explain the model's ability to generalize to new pronouns in our setting.

4.3 Consistency Results

Next, we evaluate system consistency on groups of closely related instances in WINOPRON, in order to dissect performance results and examine if systems are really right for the right reasons. We follow Ravichander et al. (2022) in operationalizing consistency by taking the score of the lowest-performing instance in the group as the group's score. We consider two groups, illustrated in Figure 4: (a) *pronoun consistency*, and (b) *disambiguation consistency*, inspired by Abdou et al.'s (2020) pair accuracy on Winograd Schemas. In both cases, we report the percentage of groups for which a model performs consistently.

Pronoun consistency measures model robustness across pronoun sets, i.e., if a model fails with even one pronoun set on a given template, then its score for that template is zero. As we consider four pronoun sets, chance is $50\%^4$, or 6 25%. Disambiguation consistency measures a system's ability to resolve a fixed pronoun to competing antecedents in paired templates. Chance is thus $0\ 5^2$, or 0 25.

SpanBERT-large is more robust to pronoun variation. As Table 3 shows, LingMess and the small and base sizes of FLAN-T5 score below chance, the former due to near-zero performance on *xe/xem/xyr*, and the latter due to poor performance overall. Interestingly, SpanBERT-large is more consistent (60.0%) than FLAN-T5-xl (55.3%) and FLAN-T5-xxl (43.9%). This indicates that despite its lower overall performance in Section 4.2, SpanBERT-large is more robust to pronominal variation.

Pronoun consistency

(a) **The counselor** informed the patient that *his* qualifications were in psychology.
(b) **The counselor** informed the patient that *her* qualifications were in psychology.
(c) **The counselor** informed the patient that *their* qualifications were in psychology.
(d) **The counselor** informed the patient that *xyr* qualifications were in psychology.

Disambiguation consistency

(a) **The counselor** informed the patient that xyr *qualifications were in psychology*.
(b) The counselor informed **the patient** that xyr *insurance covered the cost of the sessions*.

Figure 4: Example groups for scoring consistency metrics using WINOPRON templates for *counselor*, *patient* and possessive pronouns, with the antecedent bolded.

Model	PronounC	DisambigC
LingMess	*4.2*	33.3
CAW-coref	18.3	34.7
SpanBERT-base	50.0	*24.3*
SpanBERT-large	**60.0**	41.2
FLAN-T5-small	*3.9*	*0.0*
FLAN-T5-base	*0.8*	*0.0*
FLAN-T5-large	14.4	*5.4*
FLAN-T5-xl	55.3	**51.4**
FLAN-T5-xxl	43.9	43.3

Table 3: Consistency results on WINOPRON. Chance is 6.25% for pronoun consistency (PronounC) and 25% for disambiguation consistency (DisambigC). *Red, italicized numbers* are worse than chance.

The best model can only disambiguate half of the sentence pairs. Following from its high overall performance, FLAN-T5-xl has the highest disambiguation consistency score at 51.4%, just over half the template pairs we evaluate. In contrast, SpanBERT-base has disambiguation consistency below chance (24.3%). Given its reasonable overall performance, this result could stem from model bias, i.e., over-resolving a pronoun to a particular antecedent, disregarding the disambiguating context. We thus investigate bias in more detail next.

5 Pronominal Bias

So far, we have focused on coreference resolution performance and consistency and found that accusative forms and less frequent pronoun sets are harder, and models are mostly non-robust to pronominal variation and antecedent disambiguation. However, we have not established the extent to which models fail because they simply cannot perform the task, or if they are over-resolving a pronoun to a particular antecedent due to biased associations between them. Thus, we aim to disentangle performance and bias in this section.

Winogender Schemas were originally proposed to measure gender bias in coreference resolution by using pronouns (a form of lexical gender) as a proxy for social gender. Rudinger et al. (2018) then correlate incorrect resolution of English masculine and feminine pronouns with occupational statistics from the USA. By conflating lexical and social gender (see Cao and Daumé III (2021) for a critical discussion), their analysis is subject to the same limitations as their data: treating different grammatical cases of the same pronoun as equivalent, and focusing only on *he* and *she*. We thus propose a new method for evaluating pronominal bias in coreference resolution, correcting for these issues, and we then apply our method to investigate bias in SpanBERT models on WINOPRON.

5.1 Evaluating Pronominal Bias

When proposing a new method to evaluate pronominal bias in coreference resolution systems, our primary goal is to disentangle performance and bias. In other words, we should have reason to believe that the model can perform the task, and that the reason it gets an instance wrong is specifically due to bias. Additionally, we would like our method to work with an arbitrary set of pronouns of interest, and multiple surface forms of those pronouns.

Measuring Performance We first **(1)** isolate template pairs where the system attempts the task of coreference resolution as intended, i.e., the system resolves each pronoun to the occupation or participant (regardless of correctness). Next, we **(2)** focus on the template pairs that the model can *correctly* disambiguate with at least one pronoun set, p . We deem the model capable of performing coreference resolution on this set of template pairs if it can resolve them with at least one pronoun set.

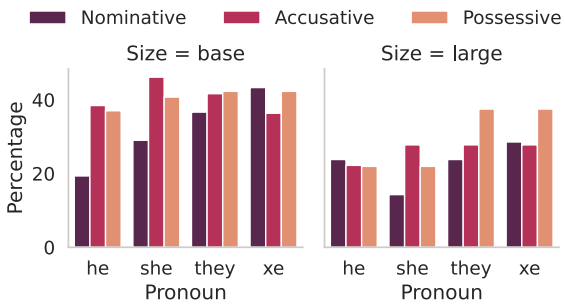

Figure 5: Percentage of model-attempted templates that show bias, for SpanBERT-base and SpanBERT-large.

Measuring Bias Of the template pairs that a model can successfully disambiguate with at least one pronoun p , we then **(3)** focus on cases where the model fails to disambiguate the exact same template pair with a different pronoun p_b p , as this is likely due to bias. If the model over-resolves p_b to the occupation, we posit that the model has a *positive bias* between p_b and that occupation. On the other hand, if it over-resolves p_b to the participant, the model is biased against associating p_b with the occupation, i.e., it has a *negative bias*.

Comparing Results With sets of positively and negatively biased occupations for each pronoun form, we want to quantify how many of a model's reasonable attempts to resolve a pronoun gave biased outputs. We thus compute the percentage of templates that result in bias (see Measuring Bias) of the total templates that a model attempts to resolve with that pronoun, given that it can correctly solve it with at least one pronoun (see Measuring Performance). This gives us a quantitative measure of "how biased" a model is which also controls for whether a model is attempting the task and can perform the task with another pronoun. In addition, we can quantify whether two models or two surface forms of a pronoun set have similar occupational biases by computing the Jaccard index (Jaccard, 1912), i.e., the size of the intersection of the biased occupation sets divided by the size of their union.

5.2 Results

We apply our method to SpanBERT-base and SpanBERT-large and collect all instances of positive and negative bias between a pronoun form and an occupation. Aggregated bias results for both models are shown in Figure 5, and Table 4 shows a sample of biased occupations for SpanBERT-large.

Pronouns	Nominative case		ccusative case		Possessive case	
	Positive	Negative	Positive	Negative	Positive	Negative
he/him/his	*engineer painter*	*receptionist secretary*	–	*dietitian secretary*	*practitioner chef*	*hairdresser secretary*
she/her/her	*hairdresser painter*	*accountant plumber*	*cashier*	*firefighter mechanic*	*practitioner painter*	*accountant surgeon*
they/them/their	–	*accountant plumber*	–	*cashier dietitian*	*advisor baker*	*accountant surgeon*
xe/xem/xyr	–	*hairdresser engineer*	–	*mechanic cashier*	*advisor baker*	*engineer supervisor*

Table 4: A sample of SpanBERT-large's biases when resolving pronouns to occupations. Positive bias: the model over-resolves the pronoun to that occupation. Negative bias: the model under-resolves the pronoun to the occupation.

Grammatical case	he	she	they	xe
Nominative	0.14	0.15	0.17	0.32
Accusative	0.12	0.10	0.25	0.29
Possessive	0.12	0.18	0.24	0.24

Table 5: Similarity of biased occupations between SpanBERT-base and SpanBERT-large, quantified with the Jaccard index (0.0 -1.0; higher is more similar).

Case pairings	he	she	they	xe
SpanBERT-base				
Nom-Acc	0.10	0.00	0.00	0.00
Acc-Poss	0.07	0.13	0.14	0.07
Nom-Poss	0.07	0.11	0.10	0.09
SpanBERT-large				
Nom-Acc	0.10	0.00	0.07	0.06
Acc-Poss	0.22	0.00	0.06	0.06
Nom-Poss	0.17	0.29	0.15	0.19

Table 6: Similarity of biased occupations across pairings of grammatical case (nom: nominative, acc: accusative, poss: possessive) of a pronoun set, quantified with the Jaccard index (0.0 -1.0; higher is more similar).

SpanBERT-base is more biased than SpanBERT-large. As Figure 5 shows, a larger percentage of SpanBERT-base's attempted and resolvable templates show biased behaviour when compared to SpanBERT-large. This pattern holds even when examining positive and negative biases separately. However, there are more negatively biased occupations than positively biased ones for both models.

Bias is qualitatively different across model sizes. In addition to being quantitatively different, we find that despite being trained and fine-tuned on the same data, there is low overlap between the occupational biases acquired by SpanBERT-base and SpanBERT-large (see Table 5). For instance, the former positively associates *she* with *machinist*, while the latter positively associates *she* with *hairdresser* and *painter*. Only *they/them/their* and *xe/xem/xyr* have slightly higher overlap, mostly due to negative bias, as these models under-resolve these particular pronouns to all occupations.

Bias does not match qualitatively across grammatical case. In other words, positive bias with *she* for an occupation does not entail positive bias with *her*. We quantify this systematically by computing Jaccard indices in Table 6, where we find that most pairings of grammatical case have very low overlap in their biases. In fact, even contradictory associations are possible; SpanBERT-base has a positive bias between *manager* and *them*, but a negative bias betweeen *manager* and *their*. Only nominative and possessive occupational biases in SpanBERT-large appear to somewhat consistently overlap with each other. Although some of these instances (e.g., negative bias for *secretary* with *he*, *him*, and *his*) align with social stereotypes (Haines et al., 2016), the overall pattern provides evidence that grammatical case in pronouns has its own set of biases that should be examined in their own right.

Bias is not additive. Even though SpanBERT-large has positive bias for *baker* and *her*, *their* as well as *xyr*, this does not imply that the model must have a negative bias between *baker* and *his*; it does not. This further highlights the need for evaluation that goes beyond binary, oppositional operationalizations of gender via pronouns.

6 Discussion

By systematically identifying and fixing issues with Winogender Schemas (Rudinger et al., 2018), we create a new dataset, WINOPRON, and find that: **(1)** different grammatical cases of pronouns show vastly different performance and bias characteristics, **(2)** pronominal biases are rich and varied, of which *he* and *she* are only the tip of the iceberg, and **(3)** model biases are complex and do not necessarily match our intuitions about them. Based on our findings, we make some recommendations for researchers who study coreference resolution and those who study bias and fairness via pronouns.

First, grammatical case is a dimension of pronominal performance and bias that warrants more study (Munro and Morrison, 2020). In particular, we hope that future work further investigates *why* accusative pronouns are harder. The patterns we demonstrate (both for performance and bias) could arise from a number of sources beyond mere frequency, including quirks of our dataset, or the distribution of semantic roles in training data for coreference resolution systems.

Second, we echo prior calls for fairness researchers to attend to the differences between social gender and terms that index it (Cao and Daumé III, 2021; Gautam et al., 2024b), to include more diversity in pronouns (Baumler and Rudinger, 2022; Lauscher et al., 2022; Hossain et al., 2023), and to move towards richer operationalizations of gender (Devinney et al., 2022; Ovalle et al., 2023) and bias (Blodgett et al., 2020). Specifically, future work on bias in coreference resolution should treat pronominal bias as distinct from (social) gender bias, defend how and why pronouns are mapped to social gender, and move beyond binary, oppositional methods of evaluation.

Lastly, as our work is a case study in how careful data curation and operationalization affects claims about system performance and bias, we emphasize the need for thoughtful data work (Sambasivan et al., 2021), and encourage the use of automatic checks when feasible, as in our work.

7 Related work

Besides Rudinger et al. (2018), there are a number of papers that tackle gender bias in coreference resolution, all of which differ from ours. Similar to Winogender Schemas, WinoBias (Zhao et al., 2018) proposes Winograd-like schemas that focus on occupations to evaluate gender bias in coreference resolution. However, WinoBias only covers *he* and *she*, rather than our coverage of all English pronoun sets by design. In addition, like Winogender, WinoBias also treats pronouns in all grammatical cases the same way. WinoNB schemas (Baumler and Rudinger, 2022) evaluate how coreference resolution systems handle singular they and plural they with similar schemas. Beyond these constructed schemas, there also exist datasets of challenging sentences found "in the wild," such as BUG (Levy et al., 2021), GAP (Webster et al., 2018), and GICOREF (Cao and Daumé III, 2021). However, as these natural datasets are not carefully constructed like Winograd-like schemas, pronouns cannot be swapped in dataset instances and still be assumed to be grammatical or coherent.

Our work is also one among several papers that investigate datasets for problems including low quality or noisy data (Elazar et al., 2024; Abela et al., 2024), artifacts (Shwartz et al., 2020; Herlihy and Rudinger, 2021; Elazar et al., 2021; Dutta Chowdhury et al., 2022), contamination (Balloccu et al., 2024; Deng et al., 2024), and issues with conceptualization and operationalization of bias (Blodgett et al., 2021; Selvam et al., 2023; Nighojkar et al., 2023; Subramonian et al., 2023; Gautam et al., 2024b). We cover many of these areas, but do not control for dataset artifacts, which we explain further in our Limitations section.

8 Conclusion

We demonstrate a number of issues with the well-known Winogender Schemas dataset, which we fix in our new, expanded WINOPRON dataset. In addition, we propose a novel way to evaluate pronominal bias in coreference resolution that goes beyond the binary and focuses on lexical gender. With our new dataset, we evaluate both supervised coreference resolution systems and language models, and find that the grammatical case of pronouns affects model performance and bias, and that bias varies widely across models, pronoun sets and grammatical cases. Our work demonstrates that measurements of bias and robustness are only as good as the datasets and metrics we use to measure them, and we call for careful attention when developing future resources for evaluating bias and coreference resolution, with attention to grammatical case, more careful operationalizations of bias, and greater diversity in the pronouns we consider.

Limitations

As in Winogender Schemas, our schemas are not "Google-proof" and could conceivably be solved with heuristics, including word co-occurrences, which is a primary concern when creating and evaluating *Winograd* schemas (Levesque et al., 2012; Amsili and Seminck, 2017; Elazar et al., 2021). The fact that we do not control for this means that our dataset gives *generous* estimates of system performance, particularly for strong language models like FLAN-T5, but it also means that this dataset is inappropriate to test "reasoning." Our dataset construction instead controls for simple system heuristics that are relevant for coreference resolution, such as always picking the first entity in the sentence, or always picking the second.

We take steps to prevent data contamination (Jacovi et al., 2023), including not releasing our data in plain text, and not evaluating with language models behind closed APIs that do not guarantee that our data will not be used to train future models (Balloccu et al., 2024). However, as we cannot guarantee a complete absence of data leakage unless we never release the dataset, we encourage caution in interpreting results on WINOPRON with models trained on data after August 2024.

Finally, we note that as our evaluation set only contains one set of templates per occupation-participant pair, our results represent a point in the distribution of bias related to that occupation. We thus echo Rudinger et al.'s (2018) view of Winogender Schemas as having "high positive predictive value and low negative predictive value" for bias. In other words, they may demonstrate evidence of pronominal bias in systems, but not prove its absence. In the case of large language models in particular, using a small number of templates for templatic evaluation is known to be brittle even to small, meaning-preserving changes to the template (Seshadri et al., 2022; Selvam et al., 2023). Our dataset's small size is a result of us requiring a tightly controlled and structured dataset to evaluate how coreference resolution varies. Thus, it may differ from realistic examples (which would have other differences that confound bias results). We wish to emphasize that in addition to controlled datasets like ours, realistic evaluation is also necessary for holistically evaluating performance, robustness and bias in coreference resolution.

cknowledgements

The authors thank Timm Dill for several rounds of patient annotation, and are grateful to Rachel Rudinger, Benjamin Van Durme, and our CRAC reviewers for their comments. Vagrant Gautam received funding from the BMBF's (German Federal Ministry of Education and Research) SLIK project under the grant 01IS22015C. Anne Lauscher's work is funded under the Excellence Strategy of the German Federal Government and States.

References

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

Kurt Abela, Kurt Micallef, Marc Tanti, and Claudia Borg. 2024. Tokenisation in machine translation does matter: The impact of different tokenisation approaches for Maltese. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 109–120, Bangkok, Thailand. Association for Computational Linguistics.

Pascal Amsili and Olga Seminck. 2017. A Google-proof collection of French Winograd schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 24–29, Valencia, Spain. Association for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtov , Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the ssociation for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Connor Baumler and Rachel Rudinger. 2022. Recognition of they/them as singular personal pronouns in coreference resolution. In *Proceedings of the 2022 Conference of the North merican Chapter of the ssociation for Computational Linguistics: Human Language Technologies*, pages 3426–3432, Seattle, United States. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th nnual Meeting of the ssociation for Computational Linguistics*, pages 5454–

5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th nnual Meeting of the ssociation for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th nnual Meeting of the ssociation for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle*. *Computational Linguistics*, 47(3):615–661.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North merican Chapter of the ssociation for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 CM Conference on Fairness, ccountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA. Association for Computing Machinery.

Karel D'Oosterlinck, Semere Kiros Bitew, Brandon Papineau, Christopher Potts, Thomas Demeester, and Chris Develder. 2023. CAW-coref: Conjunction-aware word-level coreference resolution. In *Proceedings of The Sixth Workshop on Computational Models of Reference, naphora and Coreference (CR C 2023)*, pages 8–14, Singapore. Association for Computational Linguistics.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North merican Chapter of the ssociation for Computational Linguistics: Human Language Technologies*, pages 3983–3991, Seattle, United States. Association for Computational Linguistics.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations*.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024a. Robust pronoun fidelity with english llms: Are they reasoning, repeating, or just biased?

Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. 2024b. Stop! in the name of flaws: Disentangling personal names and sociodemographic attributes in NLP. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 323–337, Bangkok, Thailand. Association for Computational Linguistics.

Elizabeth L. Haines, Kay Deaux, and Nicole Lofaro. 2016. The times they are a-changing . . . or are they not? a comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, 40(3):353–363.

Saga Hansson, Konstantinos Mavromatakis, Yvonne Adesam, Gerlof Bouma, and Dana Dannélls. 2021. The Swedish Winogender dataset. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 452–459, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Christine Herlihy and Rachel Rudinger. 2021. MedNLI is not immune: Natural language inference artifacts in the clinical domain. In *Proceedings of the 59th nnual Meeting of the ssociation for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1020–1027, Online. Association for Computational Linguistics.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st nnual Meeting of the ssociation for Computational Linguistics (Volume 1: Long Papers)*, pages

5352–5367, Toronto, Canada. Association for Computational Linguistics.

Jennifer Hu and Michael Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the ssociation for Computational Linguistics*, 8:64–77.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the ssociation for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, US* , volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Robert Munro and Alex (Carmen) Morrison. 2020. Detecting independent pronoun bias with partially-synthetic data generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2011–2017, Online. Association for Computational Linguistics.

Animesh Nighojkar, Antonio Laverghetta Jr., and John Licato. 2023. No strong feelings one way or another: Re-operationalizing neutrality in natural language inference. In *Proceedings of the 17th Linguistic nnotation Workshop (L W-XVII)*, pages 199–210, Toronto, Canada. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the ssociation for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. Association for Computational Linguistics.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 CM Conference on Fairness, ccountability, and Transparency*, pages 1246–1266.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, Brussels, Belgium. Association for Computational Linguistics.

Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 2022. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8729–8755, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 merican Chapter of the ssociation for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Nikil Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2023. The tail wagging

the dog: Dataset construction biases of social bias benchmarks. In *Proceedings of the 61st nnual Meeting of the ssociation for Computational Linguistics (Volume 2: Short Papers)*, pages 1373–1386, Toronto, Canada. Association for Computational Linguistics.

Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. *CoRR*, abs/2210.04337.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th nnual Meeting of the ssociation for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Arjun Subramonian, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett. 2023. It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. In *Findings of the ssociation for Computational Linguistics: CL 2023*, pages 3234–3279, Toronto, Canada. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the ssociation for Computational Linguistics*, 6:605–617.

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 merican Chapter of the ssociation for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

List of Occupations

The occupations along with their respective participants in parentheses are listed below in alphabetical order. This list is identical to the occupations and participants in Rudinger et al. (2018), except that we pair examiner with intern rather than victim:

accountant (taxpayer), administrator (undergraduate), advisor (advisee), appraiser (buyer), architect (student), auditor (taxpayer), baker (customer), bartender (customer), broker (client), carpenter (onlooker), cashier (customer), chef (guest), chemist (visitor), clerk (customer), counselor (patient), dietitian (client), dispatcher (bystander), doctor (patient), educator (student), electrician (homeowner), engineer (client), examiner (intern), firefighter (child), hairdresser (client), hygienist (patient), inspector (homeowner), instructor (student), investigator (witness), janitor (child), lawyer (witness), librarian (child), machinist (child), manager (customer), mechanic (customer) nurse (patient), nutritionist (patient), officer (protester), painter (customer), paralegal (client), paramedic (passenger), pathologist (victim), pharmacist (patient), physician (patient), planner (resident), plumber (homeowner), practitioner (patient), programmer (student), psychologist (patient), receptionist (visitor), salesperson (customer), scientist (undergraduate), secretary (visitor), specialist (patient), supervisor (employee), surgeon (child), teacher (student), technician (customer), therapist (teenager), veterinarian (owner), worker (pedestrian)

B nnotator Demographics

All three annotators (two authors and an additional annotator) are fluent English speakers. The two authors who create and validate templates have linguistic training at the undergraduate level. One author and one annotator have experience with using singular *they* and neopronouns, while the other author has prior exposure to singular *they* but not the neopronoun *xe*.

C nnotation Instructions

C.1 Task 1 Description

Together with this annotation protocol, you have received a link to a Google Sheet. The sheet contains 2 data columns and 2 task columns of randomized data. The data columns consist of

- Sentences which you are asked to annotate for grammaticality; and
- Questions about pronouns in the sentence, which you are asked to answer

Please be precise in your assignments and do not reorder the data. The columns have built-in data validation and we will perform further tests to check for consistent annotation.

C.1.1 Grammaticality

In the "Grammatical?" column, please enter your grammaticality judgments of the sentence, accord-

ing to Standard English. The annotation options are:

- **grammatical** (for fluent, syntactically valid and semantically plausible sentences)
- **ungrammatical** (for sentences that have any typos, grammatical issues, or if the sentence describes a situation that don't make sense, or just sounds weird)
- **not sure** (if you are not sure whether it is clearly grammatical or ungrammatical)

Examples:

- *The driver told the passenger that he could pay for the ride with cash.*
  => grammatical
- *The driver said the passenger that he could pay for the ride with cash.*
  => ungrammatical (because 'said' is intransitive in Standard English)

C.1.2 Questions about pronouns

Every sentence contains a pronoun, and the "Question" column asks whether it refers to a person mentioned in the sentence or not. The annotation options are:

- **yes** (if the pronoun refers to the person)
- **no** (if the pronoun does not refer to the person)
- **not sure** (if you are not sure about whether the pronoun refers to the person)

Examples:

- *The driver told the passenger that he could pay for the ride with cash.*
  Does the pronoun he refer to the driver?
  => no
- *The driver told the passenger that he could pay for the ride with cash.*
  Does the pronoun he refer to the passenger?
  => yes

C.2 Task 2 Description

Together with this annotation protocol, you have received a link to a Google Sheet. The sheet contains 1 randomized data column and 1 task column. Each row in the data column consists of multiple sentences, of which precisely one sentence contains a blank. Your task is to determine the appropriate pronoun to fill in the blank, and enter it in the "Pronoun" column. Here, appropriate means correct in both form and case.
The tasks are designed to be unambiguous, so please provide only one solution and do not reorder the data.

Example:

- *The driver felt unhappy because he did not make enough money. The driver wondered whether ___ should take out a loan.*
  => he

D Prompting

Table 7 shows all 10 prompt templates we use to present our task instances to FLAN-T5. Each template is presented in three variants to the model, where the options are changed:

1. No options
2. The occupation is presented first and the participant second
3. The participant is presented first and the occupation second

E dditional Results

We report additional results on double- and single-entity sentences in WINOPRON: F scores in Table 8, precision in Table 9, and recall in Table 10. Note that FLAN-T5 models generally perform worse on single-entity sentences compared to double-entity sentences because some of our prompts include options (see Section D for details) that confuse the model in this setting, despite being necessary to resolve double-entity sentences.

ID	Template
0	{t sk}\n\n{options}\nWho is {pronoun} referring to?
1	{t sk}\n\nWho is "{pronoun}" in this prior sentence (see options)?\n{options}
2	{t sk}\n\nWho is {pronoun} referring to in this sentence?\n{options}
3	Choose your answer: {t sk}\nTell me who {pronoun} is.\n{options}
4	{t sk}\nBased on this sentence, who is {pronoun}?\n\n{options}
5	Choose your answer: Who is {pronoun} in the following sentence?\n\n{t sk} \n\n{options}
6	Multi-choice problem: Which entity is {pronoun} this sentence?\n\n{t sk} \n\n{options}
7	Who is {pronoun} referring to in the following sentence?\n{t sk} \n\n{options}
8	Note that this question lists possible answers. Which person is {pronoun} referring to in the following sentence?\n{t sk} \n\n{options}
9	{t sk}\nWho is "{pronoun}"\n{options}

Table 7: Prompting templates, where "task" is filled with each dataset instance, "pronoun" is the unique third person singular pronoun in that dataset instance, and "options" are the occupation and the participant.

Data	LingMess	C W-coref	SpanBERT base	SpanBERT large	FL N-T5 small	FL N-T5 base	FL N-T5 large	FL N-T5 xl	FL N-T5 xxl
Double-entity sentences									
All	64.4	67.3	61.6	70.1	51.6	62.4	78.0	**89.0**	88.8
Nominative	73.5	77.6	67.2	77.2	51.9	65.4	85.1	94.7	**96.7**
Accusative	52.2	57.5	54.6	59.5	50.4	58.4	69.9	**82.5**	79.1
Possessive	67.4	66.5	62.9	73.6	52.3	63.4	79.1	89.7	**90.7**
he/him/his	79.2	79.6	62.8	71.5	51.5	64.1	81.5	88.8	**90.2**
she/her/her	76.3	76.6	62.1	71.6	51.5	66.1	83.3	**90.6**	89.9
they/them/their	67.5	63.7	61.2	68.9	51.8	60.5	77.0	**88.6**	88.0
xe/xem/xyr	*8.5*	*38.6*	60.4	68.5	51.4	58.7	70.3	**88.0**	87.3
Single-entity sentences									
All	73.2	75.6	**95.5**	88.0	77.3	76.3	81.5	83.1	84.3
Nominative	80.0	82.5	**99.5**	99.3	78.3	80.8	89.8	93.3	97.0
Accusative	61.1	65.0	**87.3**	67.5	76.2	69.6	69.8	70.1	66.5
Possessive	77.1	78.0	**99.8**	97.1	77.5	78.5	84.7	85.7	89.2
he/him/his	92.7	94.3	**94.7**	85.6	77.6	81.3	86.8	88.2	88.6
she/her/her	90.9	91.6	**96.2**	88.9	77.4	81.1	87.6	88.8	87.1
they/them/their	75.2	69.8	**96.0**	88.7	79.3	76.1	84.3	85.7	86.8
xe/xem/xyr	2.2	27.3	**95.2**	88.7	75.0	66.3	67.0	69.4	74.6

Table 8: F of coreference resolution systems on double- and single-entity sentences in WINOPRON. We report F overall, and split by grammatical case and pronoun set. *Red, italicized numbers* are worse than chance (50.0 for double-entity sentences and not applicable for single-entity sentences).

Data	LingMess	C	W-coref	SpanBERT base	SpanBERT large	FL	N-T5 small	N-T5 base	N-T5 large	N-T5 xl	N-T5 xxl
Double-entity sentences											
All	79.1		80.1	62.1	70.6		51.9	62.9	78.4	**89.5**	89.4
Nominative	88.3		88.7	67.4	77.4		52.1	65.7	85.4	95.1	**97.1**
Accusative	63.4		67.9	55.3	59.9		50.7	58.8	70.2	**83.2**	79.5
Possessive	86.1		83.6	63.5	74.3		52.8	64.3	79.6	90.2	**91.5**
he/him/his	79.7		80.1	63.0	71.6		51.7	64.3	81.8	89.3	**90.6**
she/her/her	77.6		77.9	62.3	71.8		51.7	66.3	83.6	**91.1**	90.3
they/them/their	79.1		80.2	61.8	69.5		52.0	60.8	77.3	**89.0**	88.5
xe/xem/xyr	**100.0**		88.1	61.3	69.3		52.1	60.1	70.7	88.6	88.0
Single-entity sentences											
All	**100.0**		**100.0**	96.0	88.4		78.9	77.6	82.4	84.0	85.6
Nominative	**100.0**		**100.0**	**100.0**	**100.0**		79.3	81.6	90.4	93.9	97.4
Accusative	**100.0**		**100.0**	88.1	67.9		77.5	70.5	70.7	71.1	68.1
Possessive	**100.0**		**100.0**	99.8	97.1		79.8	80.7	85.9	86.8	90.8
he/him/his	**100.0**		**100.0**	95.0	86.0		78.6	81.9	87.5	88.9	89.5
she/her/her	**100.0**		**100.0**	96.4	89.1		78.5	81.7	88.1	89.4	87.9
they/them/their	**100.0**		**100.0**	96.4	89.1		80.3	76.9	85.2	86.5	87.9
xe/xem/xyr	**100.0**		**100.0**	96.3	89.3		77.9	69.2	68.3	70.9	76.7

Table 9: Precision on double- and single-entity sentences overall, and split by grammatical case and pronoun set. *Red, italicized numbers* are worse than chance (50.0 for double-entity sentences, N/A for single-entity sentences).

Data	LingMess	C	W-coref	SpanBERT base	SpanBERT large	FL	N-T5 small	N-T5 base	N-T5 large	N-T5 xl	N-T5 xxl
Double-entity sentences											
All	54.2		58.0	61.1	69.7		51.3	61.9	77.7	**88.5**	88.3
Nominative	62.9		69.0	67.1	77.1		51.8	65.2	84.8	94.3	**96.3**
Accusative	*44.4*		*49.8*	54.0	59.2		50.2	58.0	69.6	**81.9**	78.6
Possessive	55.4		55.2	62.3	72.9		51.9	62.5	78.7	89.3	**90.0**
he/him/his	78.6		79.2	62.5	71.4		51.4	63.9	81.1	88.3	**89.7**
she/her/her	75.0		75.3	61.9	71.4		51.4	65.9	83.0	**90.1**	89.5
they/them/their	58.9		52.8	60.6	68.3		51.6	60.2	76.8	**88.1**	87.5
xe/xem/xyr	*4.4*		*24.7*	59.4	67.8		50.8	57.4	69.9	**87.5**	86.6
Single-entity sentences											
All	57.8		60.8	**95.1**	87.6		75.9	75.0	80.6	82.1	83.1
Nominative	66.7		70.2	**99.0**	98.5		77.3	80.0	89.2	92.7	96.6
Accusative	44.0		48.1	**86.5**	67.1		74.9	68.7	69.0	69.1	65.0
Possessive	62.7		64.0	**99.8**	97.1		75.4	76.4	83.5	84.6	87.6
he/him/his	86.4		89.2	**94.4**	85.3		76.5	80.8	86.1	87.4	87.8
she/her/her	83.3		84.4	**96.1**	88.6		76.2	80.5	87.1	88.2	86.2
they/them/their	60.3		53.6	**95.6**	88.3		78.3	75.3	83.5	85.0	85.8
xe/xem/xyr	1.1		15.8	**94.2**	88.1		72.4	63.7	65.7	68.0	72.5

Table 10: Recall on double- and single-entity sentences overall, and split by grammatical case and pronoun set. *Red, italicized numbers* are worse than chance (50.0 for double-entity sentences, N/A for single-entity sentences)