Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution

Denis Emelin[1] **Rico Sennrich**[2,1]

[1]School of Informatics, University of Edinburgh
[3]Department of Computational Linguistics, University of Zurich
D.Emelin@sms.ed.ac.uk, sennrich@cl.uzh.ch

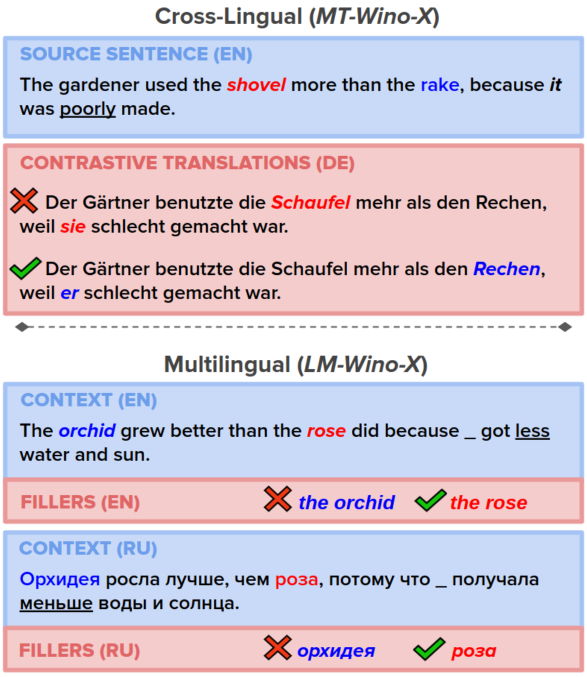
Abstract

Winograd schemas are a well-established tool for evaluating coreference resolution (CoR) and commonsense reasoning (CSR) capabilities of computational models. So far, schemas remained largely confined to English, limiting their utility in multilingual settings. This work presents *Wino-X*, a parallel dataset of German, French, and Russian schemas, aligned with their English counterparts. We use this resource to investigate whether neural machine translation (NMT) models can perform CoR that requires commonsense knowledge and whether multilingual language models (MLLMs) are capable of CSR across multiple languages. Our findings show *Wino-X* to be exceptionally challenging for NMT systems that are prone to undesireable biases and unable to detect disambiguating information. We quantify biases using established statistical methods and define ways to address both of these issues. We furthermore present evidence of active cross-lingual knowledge transfer in MLLMs, whereby fine-tuning models on English schemas yields CSR improvements in other languages.[1]

1 Introduction

Originally introduced in (Winograd, 1972), Winograd schemas (*schemas* from here on) have become an established tool for probing the ability of computational models to reason about natural language. Either viewed through the lens of coreference (CoR) as in (Levesque et al., 2012) or, more recently, framed as a gap-filling task (Sakaguchi et al., 2020), schemas are assumed to require commonsense knowledge to be resolved correctly.

Consider the following schema: *The trophy doesn't fit into the brown suitcase because **it** is too [large / small]*. Here, the pronoun ***it*** has two possible antecedents (*trophy* / *suitcase*), with the choice of the antecedent determined by the trigger word (*large* / *small*). To connect the pronoun to its true antecedent, a model must 'know' that objects that are too large cannot fit into containers and that containers that are too small cannot house objects.



Cross-Lingual (*MT-Wino-X*)

SOURCE SENTENCE (EN)

The gardener used the ***shovel*** more than the rake, because ***it*** was poorly made.

CONTRASTIVE TRANSLATIONS (DE)

✗ Der Gärtner benutzte die ***Schaufel*** mehr als den Rechen, weil ***sie*** schlecht gemacht war.

✓ Der Gärtner benutzte die Schaufel mehr als den ***Rechen***, weil ***er*** schlecht gemacht war.

Multilingual (*LM-Wino-X*)

CONTEXT (EN)

The ***orchid*** grew better than the ***rose*** did because _ got less water and sun.

FILLERS (EN) ✗ ***the orchid*** ✓ ***the rose***

CONTEXT (RU)

Орхидея росла лучше, чем роза, потому что _ получала меньше воды и солнца.

FILLERS (RU) ✗ ***орхидея*** ✓ ***роза***

Figure 1: *Wino-X* examples. *Cross-lingual* samples are used to evaluate translation models, whereas *multilingual* instances are compatible with MLLMs. Coreferent words are highlighted with the same color, while disambiguating **trigger words** are underlined.

When translating an *instance* of a schema (i.e. the schema with a fixed trigger word) into languages such as German, where pronouns and their antecedents must agree in their grammatical gender, translation models must implicitly perform the CoR step to produce accurate translations. A competent translation model is, therefore, expected to identify the correct antecedent as reflected by the target pronoun choice. In the first part of this work, we construct cross-lingual instances by aligning English instances with their translations into morphologically rich languages, so as to probe the

[1] The *Wino-X* dataset and experimental codebase can be found at https://github.com/demelin/Wino-X.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8517–8532
November 7–11, 2021. ©2021 Association for Computational Linguistics

robustness of CoR in current NMT models, as illustrated in Figure 1 (top half). In doing so, we show that models follow simplistic heuristics when attempting to resolve coreference, while failing to detect disambiguating information.

A second category of models that is expected to correctly identify coreference in multiple languages are multilingual language models. Where translation models learn to map their input to semantically equivalent sequences in the target language, MLLMs are trained on a mask-filling objective and learn to encode sentences drawn from different languages into a shared semantic space. Accordingly, schema instances correctly solved by MLLMs in one language should be equally solvable in other languages, by leveraging the same, language-agnostic representations. Similarly, improvements to model performance in one language should transfer to other languages via the shared latent space. In the second part of our work, we empirically put these assumptions to the test with multilingual schema instances, as shown in Figure 1 (bottom half), finding evidence of active commonsense knowledge transfer across languages.

Our primary contributions are as follows:

1. We introduce *Wino-X*: A dataset containing Winograd schemas in German, French, and Russian, aligned with their English analogues.

2. We benchmark the CoR performance of NMT models for each language pair, finding it to be close to chance.

3. We identify two causes underlying the poor performance of the evaluated NMT models and define ways to mitigate them.

4. We show that *Wino-X* presents a challenge for MLLMs, and observe active transfer of commonsense knowledge across languages.

2 *Wino-X*: A Contrastive Dataset of Multilingual Winograd Schemas

In order to maximize the coverage and quality of *Wino-X*, we derive multilingual schemas from *WinoGrande* (Sakaguchi et al., 2020), a large-scale, crowd-sourced corpus of English Winograd schemas. Notably, *WinoGrande* uses a *gap* token in place of an ambiguous pronoun in each schema, which can be filled by one of two preceding nouns. Based on the chosen noun, the resulting sentence either satisfies or violates commonsense constraints. Schemas are divided into two domains - social and physical. Those belonging to the former category predominantly feature names of individuals (e.g. *Mary* or *Tom*) as fillers, whereas physical samples feature objects or entities (e.g. *vase* or *cat*). Constructing cross-lingual schemas suitable for evaluating translation models requires replacing the *gap* with the ambiguous pronoun *it*, which is not possible for the social domain. Consequently, we focus our attention on the physical subset of *Wino-Grande* that contains 19,260 unique samples (9,630 schemas), with each sample representing a single instance of a monolingual, English schema.

2.1 Sample Formats

Wino-X includes samples in two formats - one for the evaluation of translation models and another for the evaluation of MLLMs. In both cases the dataset assumes a contrastive evaluation setup (Rios et al., 2017; Gardner et al., 2020), whereby evaluated models are used to rank two minimally different alternatives. Models are scored according to how frequently they rank the correct alternative above the incorrect one.

For the evaluation of NMT models, we replace the gap token with the ambiguous *it* in each sample, and pair the result with two contrastive translations. The translated *it* agrees in gender with a different antecedent in each case. For the purpose of our investigation, we focus on German, French, and Russian as morphologically rich, high-resource target languages. In the following, we refer to this set of cross-lingual samples as *MT-Wino-X*.

Evaluation of MLLMs, on the other hand, adopts the *WinoGrande* format. We translate samples without additional modifications, obtaining a set of samples for each target language that we align with their English equivalents. We refer to such multilingual samples as the *LM-Wino-X* set. Appendix A.1 provides additional examples of both formats.

2.2 From Monolingual to Multilingual

We find that not all *WinoGrande* samples are suitable for the inclusion in *Wino-X*, as replacing the *gap* with *it* can yield ungrammatical or disfluent sequences. We design a series of heuristics to filter out problematic samples, e.g. by ignoring cases where the *gap* is modified by an adjective or is part of a compound, as well as samples with animate referents. The full list is provided in Appendix A.2. We furthermore ignore samples where the *gap* is not located in the same sentence as its antecedents,

	MT-Wino-X			*LM-Wino-X*		
	EN-DE	EN-FR	EN-RU	EN-DE	EN-FR	EN-RU
# Schemas	1,887	1,499	1,119	2,917	1,396	743
# Samples	3,774	2,988	2,238	5,834	2,792	1,486

Table 1: Composition of the final *Wino-X* dataset.

	EN-DE			**EN-FR**			**EN-RU**		
Model Property	BASE	BIG	mBART	BASE	BIG	mBART	BASE	BIG	mBART
# Parameters (M)	65.5	363.5	610.9	67.7	313.1	610.9	72.5	317.9	610.9
# Training pairs (M)	39.7	538.7*	42.6	140.6	36	36.8	34.3	162*	13.9
Test BLEU	29.9	36.2	25.6	40.2	41.1	36	21.3	25.5	20.6

Table 2: Overview of the evaluated NMT models. Training size estimates were taken from corresponding publications (Ott et al., 2018; Ng et al., 2019; Tang et al., 2020). * denotes inclusion of back-translated parallel data. For mBART50, training size does not include monolingual data used in pre-training. BLEU scores were computed with SacreBLEU (Post, 2018).

to allow for a fair evaluation of models trained on sentence-level data. To reduce dataset artifacts in *Wino-X*, both instances of a schema are removed if a single one of them is filtered-out.

To obtain contrastive translations, the *gap* token is replaced with one of its fillers (which serve as the antecedents of *it*) before passing the sample through a translation engine. For all target languages, translations are obtained via the Google Translate API[2], due to its relative domain generality. Afterwards, the previously inserted filler is replaced with a pronoun of the same grammatical gender, yielding the final contrastive translation included in *MT-Wino-X*. For *LM-Wino-X* samples, the inserted filler is replaced with the *gap* token.

Following the translation step, we remove *MT-Wino-X* samples where the translated *it* has the **same gender** in both translations, resulting in an undecidable sample.[3] In contrast, for EN-FR and EN-RU portions of *LM-Wino-X*, we only remove samples where translations of both fillers have a **different gender**, as models could otherwise exploit gender agreement of verbs and adjectives to identify the correct filler.

Table 1 summarizes the primary statistics for the final dataset, with further details given in Appendix A.3. To estimate whether the constructed samples are solvable by humans, we recruited two bilingual raters for each language pair and asked them to select correct translations for a randomly drawn subset of 100 *MT-Wino-X* samples. For EN-DE, mean rater accuracy was 0.84, 0.88 for EN-FR, and 0.87 for EN-RU. Inter-rater agreement was 0.69, 0.75, and 0.77 respectively, according to Cohen's Kappa (Cohen, 1960). We replicate rater instructions in Appendix A.9. We note that since the construction of *Wino-X* relies on automated translation and linguistic analysis, the dataset is not completely free of noise. However, its impact on human performance remains within limits.

Like monolingual Winograd schemas, samples included in *Wino-X* represent particularly challenging instances of the CoR problem. However, how models handle such examples is indicative of their general language understanding capabilities. For a computational model to achieve true human parity on the translation task, it must be robust to high levels of semantic ambiguity, given that it poses little difficulty to human raters.

Next, we leverage *Wino-X* for the evaluation of coreference robustness in NMT models and of commonsense knowledge transfer in MLLMs.

3 Testing Coreference in NMT with Cross-Lingual Schemas

To probe whether NMT models can accurately identify coreference in cases requiring commonsense knowledge, contrastive translations are scored according to the sentence-level perplexity assigned to them by the evaluated model, as in Equation 1, where X is the source sequence and Y is the candidate translation:

$$PPL(Y|X) = \exp(-\frac{1}{|Y|}\sum_{i=1}^{|Y|} log_{\phi}(y_i|y_{<i}; X)) \quad (1)$$

Accuracy is based on the number of instances in which the correct translation is assigned the lower perplexity score.

[2] https://cloud.google.com/translate

[3] We use *Stanza* (Qi et al., 2020) for the linguistic analysis.

	EN-DE			EN-FR			EN-RU		
	BASE	BIG	mBART	BASE	BIG	mBART	BASE	BIG	mBART
Accuracy	0.5032	**0.5093**	0.5048	0.4960	**0.5107**	0.5030	0.4973	0.5009	**0.5049**

Table 3: Model performance on the full *MT-Wino-X* dataset. Best results per language pair are in **bold**.

	EN-DE			EN-FR			EN-RU		
Bias Type	BASE	BIG	mBART	BASE	BIG	mBART	BASE	BIG	mBART
Gender (|RBC|)	0.33	0.27	**0.37**	0.24	0.05	0.05	0.31	**0.48**	**0.44**
Positional (|RBC|)	0.16	0.17	0.14	0.05	0.07	0.15	0.07	0.06	0.05

Table 4: Model bias identified for *MT-Wino-X* samples. Higher values indicate a stronger correlation between antecedent features and model choice, and thus a greater bias. All values are statistically significant ($p < .05$). **Bold** values denote a large effect / bias size, underlined values a medium one.

3.1 Experimental Setup

Our evaluation focuses on transformer NMT models (Vaswani et al., 2017), due to their current dominance in the field. For a comprehensive examination of the relationship between model quality and CoR accuracy, we examine three model categories for each language pair: 1. **transformer-BASE** (BASE), 2. **transformer-BIG** (BIG) models distributed as part of the *fairseq* library[4], and 3. **mBART50**, a multilingual translation model built on top of a pre-trained mBART[5] (Tang et al., 2020). The inclusion of mBART50 follows the assumption that extensive pre-training may endow models with commonsense knowledge, as previously indicated for large-scale monolingual LMs (Sakaguchi et al., 2020; Bhagavatula et al., 2019; Huang et al., 2019).

BASE models are randomly initialized and trained on the concatenation of WMT news training data[6]. Data composition and pre-processing steps as well as hyper-parameter settings are summarized in Appendix A.4. As can be seen from Table 2, models differ noticeably in their size, amount of training data, and translation quality.[7]

3.2 Results and Discussion

The results of the contrastive evaluation on the full *MT-Wino-X* dataset are summarized in Table 3. All models perform at chance level (a randomly guessing model would be 50% accurate), without any observable effect of language pair, model size, training data, or monolingual pre-training.

One likely explanation is that models fall back on exploiting surface-level patterns when trying to identify the antecedent of *it*, rather than engaging in deeper language understanding. Such undesirable behaviour is facilitated by dataset biases that models are exposed to during training (Emelin et al., 2020). In their study of coreference, (Stojanovski et al., 2020) indicate that gender and positional biases can influence model behavior. To verify whether this is the case for cross-lingual Winograd schemas, we examine how strongly pronoun gender and the relative antecedent position correlates with model preference.

Importantly, in contrast to prior work, we quantify model bias explicitly as the *absolute effect size of the observed correlation* (i.e. its 'magnitude'), allowing us to directly compare between individual models and language pairs. Correlation significance is computed according to the Mann-Whitney U test (Mann and Whitney, 1947), whereas the effect size is estimated as the Rank Biserial Correlaton (RBC) score[8] (Cureton, 1956). Appendix A.5 provides additional details for both metrics.

By construction, *Wino-X* is free of gender or positional bias, since the translated *it* is guaranteed to agree with each antecedent in exactly one instance per schema, depending on the trigger word. Therefore, preferences of an unbiased NMT system should show no correlation with either property, corresponding to an $|RBC|$ score of 0. As Table 4 shows, this is not the case for the evaluated models, as we observe moderate to strong gender bias for

[4] We use single-best models in place of ensembles for the WMT19 models.

[5] We use the `mbart-large-50-one-to-many-mmt` checkpoint distributed as part of the *HuggingFace Transformers* library (Wolf et al., 2019).

[6] `http://www.statmt.org/wmt[14,20]/translation-task.html`

[7] Notably, the EN-FR BIG model had not been trained on back-translated data, unlike its EN-DE and EN-RU counterparts. We elected to tolerate this to allow for easy replication of our experiments using the same openly available, pre-trained NMT models, as well as to reduce the computational overhead and environmental impact incurred by our study.

[8] As implemented in the *pingouin* library (Vallat, 2018).

	EN-DE			EN-FR			EN-RU		
	BASE	BIG	mBART	BASE	BIG	mBART	BASE	BIG	mBART
Trigger importance	0.03	0.11*	0.12*	0.16*	0.01	0.2*	0.02	0.35*	0.08*

Table 5: Trigger importance. * denotes statistically significant differences according to paired t-tests ($p < .05$).

EN-DE and EN-RU, but not EN-FR, as well as a trivial, but statistically significant positional bias.[9]

Based on these observations, we can draw several conclusions: 1. While both bias types influence model behaviour, gender bias usually dominates positional bias, 2. Neither extensive pre-training nor multilingual training result in bias reduction for individual language pairs, and 3. The magnitude of biases in CoR is closer associated with training data properties than model properties. We verify the last point by examining the frequency with which different pronoun forms occur in the training data of our BASE models, finding that gender preferences exhibited when scoring *MT-Wino-X* mirror the pronoun gender distribution in the training data (see Appendix A.6 for relevant statistics). Surprisingly, *absolute* pronoun form frequencies appear to matter more than the likelihood of *it* being translated into a particular gender. This suggests that the frequency prior underlying the models' gender bias is surprisingly simple and, at least partly, based on raw occurrence statistics.

While model reliance on surface-level patterns provides one possible explanation for the challenging nature of *MT-Wino-X*, we also investigate whether models consider trigger terms to be especially salient when translating ambiguous pronouns.

3.3 Do Models Recognize Coreference Trigger Words?

For the estimation of salience of individual source words for the translation of *it*, we adopt the *prediction difference* (PD) technique (Li et al., 2019), shown to provide informative explanations of model behaviour by (Li et al., 2020). To apply PD to the study of coreference, we compare the probabilities assigned by the model to the correct *it* translation (w) conditioned on 1. the full source sentence (X) and 2. the source sentence without the trigger term ($X \backslash t$). To 'remove' a trigger word, its embedding is replaced with a zero vector of equal size. Salience is computed according to Equation 2, as the difference between the two probabilities.[10]

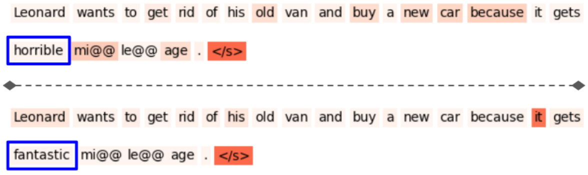

Leonard wants to get rid of his old van and buy a new car because it gets
horrible mi@@ le@@ age . </s>
Leonard wants to get rid of his old van and buy a new car because it gets
fantastic mi@@ le@@ age . </s>

Figure 2: Salience maps for two *MT-Wino-X* samples (DE side is omitted for clarity). Words that are more salient for the translation of *it* are highlighted in a deeper shade of orange. Blue frames indicate trigger words that resolve coreference ambiguity.

$$Salience(t; w, X) = P(w|X) - P(w|X \backslash t) \quad (2)$$

In order to quantify the overall *relative importance* of trigger tokens compared to non-trigger words per model, we compute **importance scores**, defined as the standardised difference between the means of salience score distributions assigned to trigger tokens and words present in both contrastive translations (i.e. non-triggers). Formally, we compute Cohen's D effect size measure, by subtracting the means of the compared distributions μ_T and μ_{NT} and dividing the result by the pooled standard deviation s, as in Equation 3. Table 5 reports the results.

$$D = \frac{\mu_T - \mu_{NT}}{s} \quad (3)$$

Across all models and language pairs, importance scores remain low[11] with the difference between salience scores lacking statistical significance in several cases. On the sentence level, this corresponds to models failing to identify trigger words required to establish coreference, as illustrated in Figure 2 for the BIG EN-DE model.

Therefore, the failure of models to perform well on the *MT-Wino-X* benchmark can be partially attributed to their inherent inability to identify information relevant for establishing coreference.

3.4 Improving CoR by Reducing Biases and Enhancing Model Awareness

Finally, we set out to improve coreference resolution in NMT models by addressing undesirable

[9] Thresholds used for interpreting the bias severity are derived in Appendix A.5.

[10] We average the salience of constituent sub-words for segmented words.

[11] Cohen's D values < 0.5 are considered to be trivial to small (Cohen, 2013).

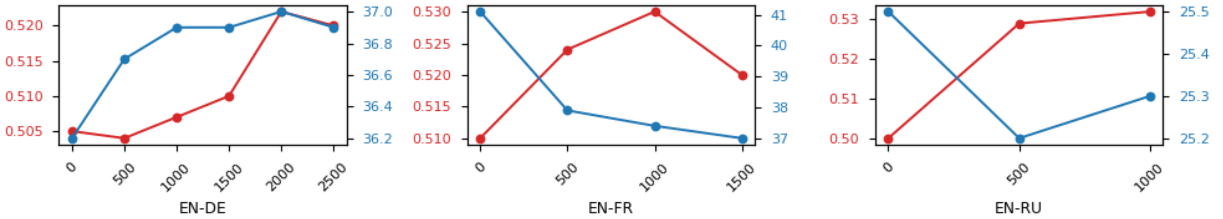


Figure 3: Few-shot fine-tuning results on *MT-Wino-X*. Red lines denote accuracy, blue lines correspond to BLEU.

biases and enhancing their ability to detect disambiguating information. Since *MT-Wino-X* is constructed to be unbiased towards antecedent gender, a straight-forward way to mitigate model bias is to fine-tune models on a fraction of the dataset, building upon the methodology proposed in (Saunders and Byrne, 2020). Given its limited size, extensive fine-tuning on *MT-Wino-X* is not feasible. However, to investigate whether bias reduction alone is sufficient to improve CoR that presupposes commonsense knowledge, we conduct a series of few-shot fine-tuning experiments.

For this purpose, we split language-specific *MT-Wino-X* datasets into training, development, and test sets, taking care that both instances belonging to the same schema are assigned to the same split. For all experiments, development and test sets are fixed, containing 200 and 1k samples, respectively. Training set size is varied in increments of 500 up to 2k for EN-DE, 1.5k for EN-FR, and 1k for EN-RU. All models are fine-tuned until convergence as determined by early-stopping, with hyper-paremeter settings discussed in Appendix A.7. We focus on the BIG models, measuring the effect of increased training size on accuracy and translation quality.

As shown in Figure 3, fine-tuning yields slight improvements in accuracy for all language pairs, up to 3.2% for EN-RU. In parallel, we observe a substantial reduction in gender bias in fine-tuned models, using the methodology from §3.2. Exposing translation models to 2.5k samples for EN-DE and 1k for EN-RU reduces gender bias by **71%** and **73%**, respectively, from 0.24 to 0.07 and from 0.49 to 0.13.[12] **Still, debiasing alone is not sufficient to substantially increase CoR accuracy**.

We also note that fine-tuning has a mixed effect on test BLEU which increases for EN-DE but degrades for EN-FR and, to a lesser extent, EN-RU. An analysis of EN-DE test translations before and after fine-tuning shows an increased pronoun coverage for the fine-tuned model, with most pronounced improvements detected for masculine and feminine pronoun forms (Table 6), corroborating the quantitative reduction in gender bias.

Source	Feminine	Masculine	Neutral
Reference	340	476	380
Pre-trained	270	410	321
Fine-tuned	**290**	**420**	**326**

Table 6: Pronoun frequencies in BIG EN-DE translations, compared to the *newstest2020* reference.

Since bias reduction alone does not suffice to address the unique challenges presented by *MT-Wino-X*, we additionally experiment with equipping translation models with an inductive bias that facilitates accurate pronoun translation. To accomplish this, we define the *Pronoun Penalty* (PP) objective that actively penalizes translation models for assigning higher probability to an incorrect pronoun form during training.[13], so as to encourage models to better utilize trigger words. The objective is defined in Equation 4, where CE is the smoothed cross-entropy loss, λ is the scaling factor, $r \in R$ are correct target pronouns found in the reference translation, and $a \in A$ are alternative, incorrect pronoun forms for each correct pronoun (e.g. [*er*, *es*] if the correct German pronoun is *sie*).

$$L(S) = CE(S) + \lambda \sum_{r=1}^{|R|} PP(r) \tag{4}$$

$$PP(r) = 1 - \frac{P(r)}{N} + \frac{\max_{a \in A} P(a)}{N} \tag{5}$$

$$N = P(r) + \sum_{i=1}^{|A|} P(a_i) \tag{6}$$

We fine-tune the BIG models on the largest training set for each language pair with this enhanced objective, and present the results in Table 7.[14] The new objective substantially improves accuracy for

[12] Initial gender biase values (i.e. 0.24 and 0.49) are recomputed on test sets used in the few-shot experiments. Given the low initial gender bias in EN-FR BIG (0.024), fine-tuning has no noticeable effect.

[13] For simplicity, we only consider singular pronouns in the nominative case, e.g. [*er*, *sie*, *es*] for DE.

[14] $\lambda = 100$ for all language pairs.

EN-DE and EN-FR, by 4-7%, while no noticeable difference can be observed for EN-RU. Crucially, the observed improvements correlate with an increase in trigger word importance. Reusing the method introduced in §3.3, we find trigger importance increase by a factor of **1.5** for EN-DE and **4.25** for EN-FR compared to models fine-tuned without PP, from 0.12 to 0.18 and 0.04 to 0.17.[15]

Regime	EN-DE	EN-FR	EN-RU
Pre-trained	0.51 (36.2)	0.51 (41.1)	0.5 (25.5)
Fine-tuned	0.52 (36.9)	0.52 (37)	**0.53** (25.3)
+ PP	**0.56** (36.6)	**0.59** (39.4)	**0.53** (25.3)

Table 7: *MT-Wino-X* accuracy of models with different training regimes. Test BLEU in parentheses.

Overall, our findings indicate that coreference remains an unsolved challenge in machine translation, especially in cases requiring commonsense knowledge. **While debiasing models leads to improved CoR accuracy, inductive biases that enable models to detect disambiguating information can be more important still.**

4 Testing Cross-Lingual Transfer in MLLMs

Having thus probed the capacity and limitations of NMT models for solving cross-lingual *Wino-X* samples, we now turn to MLLMs.

4.1 Experimental Setup

Our investigation seeks to answer two questions: 1. *To what extent can MLLMs solve Winograd schemas in different languages?* and 2. *Does commonsense knowledge actively transfer across languages?* Should the latter be the case, it could substantially reduce the need for language-specific commonsense knowledge bases that usually require significant human effort to construct and expand (Speer et al., 2017). Our experiments focus on the XLM-RoBERTa (XLM-R) model introduced in (Conneau et al., 2020). Structurally similar to the decoder of a transformer NMT model, XLM-R is trained on monolingual as well as parallel data covering 100 diverse languages, to induce language-agnostic representations in a shared semantic space. Intuitively, sharing representations across languages should facilitate commonsense knowledge transfer, although it is yet unclear to what extent this holds true for Winograd schemas.

Analogous to our evaluation of NMT models, MLLMs are examined in the contrastive setting. As input, models receive a schema instance containing a *gap*, as depicted in Figure 1 (bottom half), which is replaced with a model-specific <*MASK*> token used during pre-training. Conditioned on this input, we compute sentence-level pseudo-perplexities (PPPL) (Salazar et al., 2020) for two completions of the input sequence, each with a different filler that replaces the <*MASK*> token. The completion assigned the lowest PPPL indicates the model's preference towards a specific gap-filler, which informs model accuracy.

4.2 Results

As a first step, we measure the zero-shot performance of XLM-R BASE (∼270M parameters) and LARGE (∼550M parameters) models[16] on the full *LM-Wino-X* datasets, summarizing the results in Table 8. Accuracy remains comparatively low across the board, with the BASE model scoring close to chance level. On the other hand, the XLM-R LARGE variant substantially outperforms its BASE analogue and demonstrates roughly comparable performance across all examined languages.

	EN-DE		**EN-FR**		**EN-RU**	
	EN	DE	EN	FR	EN	RU
BASE	0.53	0.53	0.54	0.53	0.52	0.52
LARGE	**0.62**	**0.61**	**0.63**	**0.6**	**0.62**	**0.59**

Table 8: XLM-R accuracy on *LM-Wino-X*. Since dataset composition and size differs between language pairs as detailed in §2.2, for EN-X, EN denotes model performance on the EN side of the pair-specific dataset, and X on the aligned non-EN language.

4.3 Is Monolingual Data Enough for Multilingual CSR?

Of central interest to our investigation is whether fine-tuning models on schema instances in a *primary language*, e.g. EN, also improves CSR in a *transfer language*, e.g. DE, and how this improvement compares to directly fine-tuning the model on the latter. We conduct a series of few-shot experiments to answer this question, while exploring the relationship between cross-lingual commonsense knowledge transfer and the amount of fine-tuning data. Due to its greater efficiency, our investiga-

[15] As with bias values, initial trigger importance scores are re-computed on test sets used in few-shot experiments. Fine-tuning has a limited effect on EN-RU which had the highest initial importance scores.

[16] We use the *HuggingFace Transformers* library.

	EN-DE		EN-FR		EN-RU	
	EN	DE	EN	FR	EN	RU
Accuracy FT	0.67	0.60	0.67	0.59	0.65	0.57
Accuracy Δ	14%	7%	13%	6%	13%	5%

Table 9: Test accuracy of XLM-R BASE fine-tuned on *WinoGrande*. DE, FR, and RU are transfer languages not seen during fine-tuning. Δ denotes the accuracy increase compared to the first row of Table 8.

tion is focused on XLM-R BASE[17]. Analogous to experiments in §3.4, we split the *LM-Wino-X* data into training, development, and test sets, keeping development and test sizes fixed at 200 and 1k samples, while varying the size of the training set in increments of 500. Instances derived from the same schema are assigned to the same set.

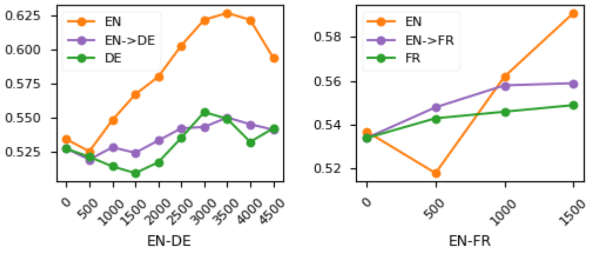


Figure 4: Few-shot fine-tuning results on *LM-Wino-X*. EN->X denotes zero-shot knowledge transfer to language X after training the model on EN samples only.

To adopt XLM-R to the studied task, it is fine-tuned on target sequences containing the correct gap-filler with the masked language modeling objective. Models are trained until convergence as determined by early-stopping, with hyper-parameters given in Appendix A.8. We treat EN as the primary language and evaluate knowledge transfer toward DE and FR[18], summarizing the results in Figure 4. Improved accuracy is observed for all models. However, fine-tuning benefits EN models most as the amount of training samples increases, which may be linked to EN being the dominant language in the XLM-R pre-training corpus (Conneau et al., 2020). More importantly, we can observe a substantial transfer of commonsense knowledge between languages. Models fine-tuned on EN and evaluated on DE / FR often achieve higher accuracy than models directly fine-tuned on the transfer language.

To shed light on commonsense knowledge transfer beyond the few-shot setting, we additionally fine-tune instances of XLM-R on the entirety of WinoGrande and evaluate them on the few-shot test sets.[19] As can be seen from Table 9, commonsense knowledge transfer benefits from the increase in training data, with improvements in the transfer languages being roughly half of those observed for the primary language. **This indicates that large-scale, monolingual commonsense resources can significantly contribute towards building models capable of CSR in a wide variety of languages**.

5 Related Work

Winograd schemas have been widely adopted in recent years for the study of pronominal coreference and CSR (Kocijan et al., 2020). Several datasets have been proposed, differing in whether schemas are authored by experts (Levesque et al., 2012; Wang et al., 2019) or composed by crowd-workers (Isaak and Michael, 2019; Sakaguchi et al., 2020). Crucially, the majority of such resources is in English, with the notable exception of (Amsili and Seminck, 2017; Melo et al., 2019; Bernard and Han, 2020) (each contain a few hundred examples). The process by which we extend monolingual schemas into other languages shares similarities with (Stanovsky et al., 2019), while also modifying the English schemas and incorporating a more sophisticated set of filtering heuristics, due to differences in the examined tasks.

Similarly, the study of coreference has a long tradition in machine translation. Several CoR datasets have been proposed in the past, including (Guillou and Hardmeier, 2016; Bawden et al., 2018; Müller et al., 2018; Stojanovski et al., 2020). Among those, that of (Stojanovski et al., 2020) is most relevant to our work. While it contains samples that require world knowledge to resolve coreference, they are constructed from a fixed set of templates and remain limited to EN-DE. In contrast, *Wino-X* encompasses multiple target languages, while offering greater linguistic and thematic diversity.

Finally, while cross-lingual transfer in MLLMs has received much attention in the past (Conneau et al., 2018, 2020; Hu et al., 2020; Liang et al., 2020), research on CSR in multiple languages remains limited, with (He et al., 2020) being the only relevant machine translation study known to us. Concurrent to our work, (Lin et al., 2021) examine whether MLLMs can perform multilingual CSR on tasks unrelated to Winograd schemas.

[17] We were unable to train XLM-R LARGE as our hardware could not accommodate its significant size outside inference.

[18] Due to its limited size, EN-RU data is excluded from the few-shot evaluation.

[19] Excluding samples found in each test set from training.

6 Conclusion and Outlook

In this work, we introduced *Wino-X*, a dataset containing cross-lingual and multilingual Winograd schemas. Based on this resource, we showed that NMT models struggle to correctly resolve coreference that presupposes commonsense knowledge, due to over-reliance on dataset artifacts and general inability to detect disambiguating information. We defined methods to quantify biases and trigger word importance in a principled way, and proposed strategies for reducing the former while increasing the latter. For MLLMs, we presented evidence of commonsense knowledge transfer, showing that transferring knowledge from English to another language can lead to similar (or greater) improvements as directly fine-tuning on transfer languages. Overall, our study identifies existing difficulties in cross-lingual CoR and CSR, discusses potential causes, and offers initial ways to mitigate them.

In future work, we intend to further improve the handling of coreference in NMT by reducing undesirable biases and introducing useful ones. For MLLMs, future efforts can be directed towards identifying categories of knowledge that do not benefit from cross-lingual transfer, to effectively guide data collection in lower-resourced languages.

Acknowledgments

We thank Sabine Weber for her valuable comments on the topic of Winograd schema, our data evaluators for their contributions to this work, as well as the anonymous reviewers for their helpful feedback. Rico Sennrich acknowledges support of the Swiss National Science Foundation (MUTAMUR; no. 176727).

Ethical Considerations

Since our work introduces a novel resource, we include a Data Statement (Bender and Friedman, 2018) as a concise overview of its provenance and construction. We hope this will motivate the research community to adopt the dataset for projects relating to cross-lingual natural language understanding by increasing transparency.

A. CURATION RATIONALE: We discuss the filtering criteria applied to *WinoGrande* samples and their translations in §2.2 and §A.2. In enforcing conservative selection criteria, our aim is to ensure grammaticality of the semi-automatically constructed samples and to minimize the percentage of undecidable or disfluent instances.

B. LANGUAGE VARIETY: The collected dataset contains English, German, French, and Russian sentences. English sentences were authored by human crowd-workers, while translations into other languages were obtained from an online translation service. Since (Sakaguchi et al., 2020) do not provide demographics of workers involved in data collection, we cannot report on the dominant variety of English. Due to their origin, translations into DE, FR, and RU are likely to exhibit features of neural translationese (Graham et al., 2020).

C. SPEAKER DEMOGRAPHIC: N/A

D. ANNOTATOR DEMOGRAPHIC: We appropriate this section to summarize the demographics of raters involved in evaluating the dataset quality, as detailed in §2.2. Of the 6 annotators involved (two per language pair), all were bilingual speakers with native or native-like proficiency in both English and German / French / Russian. All six were of European origin, between 25-35 years of age, and held a graduate degree. Four of the raters identified as female and two as male.

E. SPEECH SITUATION: The dataset was constructed semi-automatically using scripts distributed in the project's repository. Raters submitted their judgments in the course of a single week and had the opportunity to contact the primary author with clarifying questions.

F. TEXT CHARACTERISTICS: *Wino-X* contains a collection of cross-lingual and multilingual Winograd schemas for the study of coreference resolution and commonsense reasoning in NMT models and MLLMs. Due to the relative simplicity of scenarios described by the schemas, it is highly unlikely for the dataset to have significant ethical implications.

G. RECORDING QUALITY: N/A

H. OTHER: N/A

I. PROVENANCE APPENDIX: According to (Sakaguchi et al., 2020), *WinoGrande* was collected through the Amazon Mechanical Turk (AMT) platform. Workers had to meet a minimum qualification that required 99% approval rate and 5k AMT approvals in total. For composing twin sentences corresponding to a single schema, workers were awarded \$0.4. Each collected sample was subsequently validated by three other crowd-workers, with 68% of samples deemed to be valid. For each sentence validation, workers were reimbursed with \$0.03. See (Sakaguchi et al., 2020) for a more extensive discussion of *WinoGrande*.

References

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.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Timothée Bernard and Ting Han. 2020. Mandarinograd: A chinese collection of winograd schemas. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 21–26.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Edward E Cureton. 1956. Rank-biserial correlation. *Psychometrika*, 21(3):287–290.

Denis Emelin, Ivan Titov, and Rico Sennrich. 2020. Towards detecting and exploiting disambiguation biases in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7635–7653.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1307–1323.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81.

Liane Guillou and Christian Hardmeier. 2016. Protest: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643.

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3662–3672.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401.

Nicos Isaak and Loizos Michael. 2019. Winoflexi: A crowdsourcing platform for the development of winograd schemas. In *Australasian Joint Conference on Artificial Intelligence*, pages 289–302. Springer.

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

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

Jierui Li, Lemao Liu, Huayang Li, Guanlin Li, Guoping Huang, and Shuming Shi. 2020. Evaluating explanation methods for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 365–375.

Xintong Li, Guanlin Li, Lemao Liu, Max Meng, and Shuming Shi. 2019. On the word alignment from neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1293–1303.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for common sense reasoning. In *To appear.*

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Robert E McGrath and Gregory J Meyer. 2006. When effect sizes disagree: the case of r and d. *Psychological methods*, 11(4):386.

Gabriela Melo, Vinicius Imaizumi, and Fábio Cozman. 2019. Winograd schemas in portuguese. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 787–798. SBC.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *WMT*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Annette Rios, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. *arXiv preprint arXiv:2004.04498*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

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

Dario Stojanovski, Benno Krojer, Denis Peskov, and Alexander Fraser. 2020. Contracat: Contrastive coreference analytical templates for machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4732–4749.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Raphael Vallat. 2018. Pingouin: statistics in python. *The Journal of Open Source Software*, 3(31):1026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

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*.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Supplementary Material

A.1 Additional Wino-X Examples

Additional *MT-Wino-X* examples are provided in Table 10, while Table 11 contains further *LM-Wino-X* entries.

A.2 Filtering Heuristics

To obtain grammatical sentences after replacing the *gap* token with *it*, we exclude *WinoGrande* samples from *Wino-X* if:

- Either referent is animate (e.g. *teacher*, *baker*)
- The *gap* token is part of a compound noun or a noun phrase
- Either referent is a plural noun
- The gap token is modified by an adjective

To improve the quality of our constructed cross-lingual and multilingual schemas, we aim to reduce potential sources of noise by furthermore excluding samples if:

- The translated *it* or gap-filler is not in the nominative case
- Either antecedent denotes an activity (e.g. *singing* or *playing the piano*) (due to issues it presents to morphological analyzers)

Additionally, we use a grammar checker[20] to ensure that the insertion of *it* does not introduce grammatical errors.

A.3 Additional Dataset Statistics

Table 12 summarizes the fine-grained statistics for the *MT-Wino-X* and *LM-Wino-X* datasets.

A.4 NMT Training Details

EN-DE and EN-RU models are trained on the concatenation of WMT20 news task data, with *newstest2019* used for development and *newstest2020* serving as the text set. For EN-DE, we exclude the *Wiki Titles v2* corpus. The EN-FR model, on the other hand, is trained on the WMT14 news task data, augmented with *ParaCrawl v8*[21]. We use *newstest2013* as the development set and test on *newstest2014*. All data is cleaned by removing sentence pairs with a source-to-target length ratio exceeding 2 or identified as belonging to unrelated languages by *langid*[22]. We tokenize all datasets using *Moses* scripts[23] and employ the *subword-nmt* library[24] (Sennrich et al., 2016) to segment words. Subword segmentation used 32k merge operations and a vocabulary threshold of 50.

Hyper-parameter settings are provided in Table 13. We adopt the same settings for all three models. The only exception is the use of tied embeddings for EN-DE and EN-FR, but not EN-RU, as recommended in (Ng et al., 2019). Parameters specific to the transformer architecture (e.g. layer size, number of attention heads) correspond to the BASE configuration in (Vaswani et al., 2017). Other hyper-parameters not covered in Table 13 use the default *fairseq* settings for the 'transformer' architecture. All models were trained on NVIDIA RTX 2080 Ti cards until convergence according to early stopping ($\sim$20 hours each).

Hyper-parameter	**Value**
LR	7e-4
LR schedule	*inverse_sqrt*
Batch size	4,096 tokens
# Gradient accumulation steps	6
Optimizer	Adam
Adam betas	0.9, 0.98
Dropout *p*	0.1
Warm-up updates	4k
Max # Epochs	1k
Validation frequency	5k updates
Early stopping patience	3
Random seed	42

Table 13: Hyper-parameters for training **BASE** models.

A.5 Statistical Methods

To estimate the statistical significance of the correlation between the gender of the translated *it* and model preference, the Mann-Whitney U test combines translations preferred by the model (i.e. those assigned the lower PPL) and those rejected by the model and ranks them according to the numerical ID that corresponds to the gender of the *it* translation (i.e. 1=*masculine*, 2=*feminine*, 3=*neutral*). Subsequently, the U-value is computed according to Equations 7-9, where R_1 denotes the sum of ranks of translations preferred by the model and n_1 their total count, while R_2 denotes the sum of ranks of translations rejected by the model and n_2

[20] LanguageTool for python.

[21] https://paracrawl.eu/

[22] https://github.com/saffsd/langid.py

[23] https://github.com/moses-smt/mosesdecoder

[24] https://github.com/rsennrich/subword-nmt

Dataset	Sample
EN-DE	**Source Sentence**: I dusted the **dresser** in the bedroom with a **rag** until **it** was free of dust. **Correct Translation**: Ich staubte die **Kommode** im Schlafzimmer mit einem Lappen ab, bis **sie** staubfrei war. **Incorrect Translation**: Ich staubte die Kommode im Schlafzimmer mit einem **Lappen** ab, bis **er** staubfrei war.
EN-FR	**Source Sentence**: Stacey used the company **credit card** to buy a **plane ticket**, but **it** was declined. **Correct Translation**: Stacey a utilisé **la carte de crédit** de l’ entreprise pour acheter un billet d’ avion, mais **elle** a été refusée. **Incorrect Translation**: Stacey a utilisé la carte de crédit de l’ entreprise pour acheter un **billet d’ avion**, mais **il** a été refusé.
EN-RU	**Source Sentence**: Dana could not hang the **artwork** on her **wall** because **it** was too thin. **Correct Translation**: Дана не могла повесить произведение искусства на **стену**, потому что **она** была слишком тонкой. **Incorrect Translation**: Дана не могла повесить **произведение искусства** на стену, потому что **оно** было слишком тонким.

Table 10: Additional *MT-Wino-X* examples. Highlighting signifies coreference.

Dataset	Sample
EN-DE	**EN Context**: Adam chose to sleep on a **sofa** instead of a **bed** because _ was much more comfortable. **Correct Filler**: **the sofa** **Incorrect Filler**: **the bed** **DE Context**: Adam entschied sich dafür, auf einem **Sofa** statt auf einem **Bett** zu schlafen, weil _ viel bequemer war. **Correct Filler**: **das Sofa** **Incorrect Filler**: **das Bett**
EN-FR	**EN Context**: The bartender poured the juice from the **blender** into the **cocktail glass** until _ was full. **Correct Filler**: **the glass** **Incorrect Filler**: **the blender** **FR Context**: Le barman versa le jus du **mixeur** dans le **verre** à cocktail jusqu’à ce que _ soit plein. **Correct Filler**: **le verre** **Incorrect Filler**: **le mixeur**
EN-RU	**EN Context:** The man took off the **tank top** and put on the **t-shirt**, because _ was sweaty. **Correct Filler:** **the tank top** **Incorrect Filler:** **the t-shirt** **RU Context:** Мужчина снял **майку** и надел **футболку**, потому что _ была потной. **Correct Filler:** **майка** **Incorrect Filler:** **футболка**

Table 11: Additional *LM-Wino-X* examples. Highlighting signifies coreference.

	MT-Wino-X		***LM-Wino-X***			
	Mean Sentence Length	Mean Translation Length	Mean EN Context Length	Mean X Context Length	Mean EN Filler Length	Mean X Filler Length
EN-DE	17.8 (2.86)	17.15 (3.1)	17.84 (2.86)	17.16 (3.11)	2.04 (0.19)	2 (0.0)
EN-FR	17.85 (2.9)	20 (3.87)	18.01 (2.86)	20.24 (3.74)	2.02 (0.13)	2 (0.0)
EN-RU	17.73 (2.87)	14.86 (2.99)	18.06 (2.97)	15.34 (3.07)	2.02 (0.14)	2 (0.0)

Table 12: Additional dataset statistics. *X* stands for the language aligned with En for each language pair. Length is computed in tokens based on Moses-tokenized sentences. Values in parentheses denote standard deviation.

their respective total count.

$$U = min(U_1, U_2) \tag{7}$$

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \tag{8}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} \tag{9}$$

To obtain the p-values, U-values are subjected to tie correction and normal approximation. Significance of the positional bias is computed following the same procedure, with ranking taking place according to the relative antecedent location.

In order to compute the RBC values, test sentences are divided into two groups - one containing translations that are preferred by the model and another comprised of the rejected translations. Next, all possible pairs are constructed between the two groups, pairing together each translation from one group with all translations in the other. The proportion of pairs f where the pronoun ID of the preferred translation is greater than that of the rejected translation is computed, as well as the proportion of pairs u where the opposite relation holds. The

	EN-DE			**EN-FR**		**EN-RU**		
Model Property	Masc.	Fem.	Neut.	Masc.	Fem.	Masc.	Fem.	Neut.
# Preferred by model	982	985	1,807	1,893	1,095	1,674	311	124
# ref *it* translations	175,552	473,880	2,113,923	4,719,590	1,460,420	236,422	225,560	102,189
# absolute ref. occurrences	1,757,265	-*	4,504,505	21,715,292	6,026,101	1,035,361	508,117	165,369

Table 14: Pronoun frequencies in *MT-Wino-X* translations preferred by BASE models and found in the training data. *The German *sie* is highly polysemous and, as such, not included in the absolute counts, since disambiguation via linguistic analysis of ∼10M candidate sentences (e.g. with *Stanza*) was computationally prohibitive.

RBC value is obtained according to Eqn.10.

$$RBC = f - u \tag{10}$$

As we are only interested in the effect size and not in the direction of the effect, we take its absolute value to signify bias strength. Positional bias is estimated in the same manner.

A common practice for interpreting effect size strength is the adoption of Cohen's benchmark (Cohen, 2013), which posits that the effect size d is large if $d >= 0.8$, medium if $d >= 0.5$, and small if $d >= 0.2$. It is, however, not inherently applicable to the interpretation of RBC, due to its insensitivity to the *base rate* - the size ratio between the two groups denoted by the dichotomous variable, i.e. whether a translation is preferred or rejected by the model. For a detailed discussion, see (McGrath and Meyer, 2006). To apply the aforementioned thresholds to RBC, we use the conversion formula in Equation 11 (McGrath and Meyer, 2006), where $p1$ and $p2$ represent the proportions of groups described by the dichotomous variable, with $p_1 = p_1 = 0.5$. Within the contrastive evaluation setting, the base rate is guaranteed to equal 1, since for each sample, one translation will be preferred by the model while the other one is rejected.

$$threshold = \frac{d}{\sqrt{d^2 + \frac{1}{p_1 p_2}}} \tag{11}$$

The adjusted effect size thresholds are, therefore, as follows: *small* if $d >= 0.1$, *medium* if $d >= 0.24$, and *large* if $d >= 0.37$.

A.6 Pronoun Frequencies

For EN-DE, our BASE model strongly favours neutral antecedents, preferring them over the alternative in ∼48% of samples, while they represent the correct choice in just ∼31% of the dataset. Looking at the training data, we find that translations of *it* are 4.5-12 times more likely to have the neutral gender than female and male, respectively. A similar trend can be observed for EN-FR, where *it* is translated as male in ∼63% of samples favoured by the model (which is correct in ∼50% of the dataset), with translations into the male gender being 3.2 times more likely than female in the training data. Male gender is even more dominant for EN-RU, where it is preferred by the model in ∼79% of instances (and correct in just ∼40% of the dataset).

Importantly, the likelihood of *it* being translated as male or female in the EN-RU training data is roughly equal, with translation into male being 1.05 times more likely, yet the absolute frequency of the male pronoun is roughly twice as high compared to the female form. A similar picture emerges for the EN-FR data, where the male pronoun is 3.6 times more frequent than its female analogue, overall. It is difficult to estimate the absolute frequency of the German female pronoun, as it is highly polysemous. Table 14 summarizes the corresponding statistics.

A.7 NMT Fine-Tuning

To fine-tune the BASE and BIG NMT models, we use the same settings as provided in §A.4, but set the learning rate to 1e-7, reduce the total batch size to 8 sentence pairs, and forego any warm-up steps. Models are fine-tuned to convergence according to early-stopping, with patience set to 3 validation steps. Validation takes place after each completed training epoch. The optimal LR was determined via grid search over [1e-5, 1e-6, 1e-7].

Settings for fine-tuning mBART are summarized in table 15. Hyper-parameters not covered use the default setting in *HuggingFace Transformers*.

Hyper-parameter	**Value**
LR	1e-5
# Gradient accumulation steps	1
Batch size	16 sentence pairs
Max # Epochs	1k
Validation frequency	1 epoch
Early stopping patience	3
Random seed	42

Table 15: Settings used to fine-tune **mBART50**.

A.8 MLLM Fine-Tuning

We provide the fine-tuning hyper-parameters used in conjunction with XLM-R BASE and LARGE in Table 16. As before, setting not covered in the table correspond to their default value in *Transformers*. Same settings are used for all language pairs. The optimal LR was determined via grid search over [1e-5, 1e-6, 1e-7].

Hyper-parameter	Value
LR	1e-7
# Gradient accumulation steps	1
Batch size	16 sentence pairs
Max # Epochs	1k
Validation frequency	1 epoch
Early stopping patience	3
Random seed	42

Table 16: Settings used to fine-tune **XLM-R**.

A.9 Rater Instructions

Once you open the form you were given a link to, you will see a sheet containing ∼100 rows, with each row representing an individual sample for you to annotate. Each row is subdivided into 4 fields: SENTENCE, TRANSLATION_1, TRANSLATION_2, and WHICH TRANSLATION IS BETTER?

Please begin the annotation of each row by first reading the sentence given in the SENTENCE field. Each SENTENCE should contain the English pronoun "it" as well as several nouns. One of the nouns should be identifiable as the referent of "it", i.e. as denoting the object or entity that "it" clearly refers to. For instance, given the SENTENCE "The trophy does not fit into the suitcase because ***it*** *is too small", the bolded* ***it*** *clearly refers to suitcase rather than trophy, since a suitcase can be too small to fit a trophy, but a trophy cannot be too small to fit inside a suitcase.*

TRANSLATION_1 and TRANSLATION_2 provide two alternative, minimally different translations of SENTENCE. The primary difference between both translations is the gender of the pronoun representing the translation of the ambiguous "it" in SENTENCE. Continuing with our running example, TRANSLATION_1 could be "Die Trophäe passt nicht in den Koffer, weil er zu klein ist", while TRANSLATION_2 could be "Die Trophäe passt nicht in den Koffer, weil sie zu klein ist". In TRANSLATION_1, "it" has been translated as the German pronoun er that unambiguously refers to Koffer (corresponding to the English "suitcase"), as both are masculine in gender. On the other hand, in TRANSLATION_2, "it" is translated as the German pronoun sie that unambiguously refers to Trophäe (corresponding to the English "trophy"), as both are feminine in gender. Given that things cannot usually be too small to fit into receptacles, TRANSLATION_1 should be judged as correct, rather than TRANSLATION_2.

When annotating each example, please select the most appropriate option from the drop-down menu in the WHICH TRANSLATION IS BETTER? column. If you think that TRANSLATION_1 is accurate or have a preference towards it (e.g. based on your world knowledge / common sense), please choose "1". If you think that TRANSLATION_2 is accurate or have a preference towards it, please choose "2". If both translations are perfectly equally likely, please choose "BOTH". If the translation quality is insufficient for you to make a confident judgment, please select "BAD SAMPLE".

Since the translations were machine-generated, we ask you to be lenient towards translation errors that do not affect the pronoun disambiguation. If the translation is not perfect, e.g. containing odd structure or mistranslated words, but you're still able to identify the correct pronoun translation, please indicate your translation choice, rather than marking the sample as bad.

TRANSLATION_1 and TRANSLATION_2 will always differ as to how "it" is translated, but may have other surface-level differences, as well. As long as both translations convey similar content, we encourage you to ignore any differences other than the translation of "it" for the purpose of your judgments.