Uncertainty Quantification for Evaluating Gender Bias in Machine Translation

Anonymous ACL submission

Abstract

The predictive uncertainty of machine translation (MT) models is typically used as a quality estimation proxy. In this work, we posit that apart from confidently translating when a single correct translation exists, models should also maintain uncertainty when the input is ambiguous. We use uncertainty to measure gender bias in MT systems. When the source sentence includes a lexeme whose gender is not overtly marked, but whose target-language equivalent requires gender specification, the model must infer the appropriate gender from the context and can be susceptible to biases. Prior work measured bias via gender accuracy, however it cannot be applied to ambiguous cases. Using semantic uncertainty, we are able to assess bias when translating both ambiguous and unambiguous source sentences, and find that high translation accuracy does not correlate with exhibiting uncertainty appropriately, and that debiasing affects the two cases differently.[1]

1 Introduction

Language is inherently ambiguous, and meaning is often resolved through context. However, not all ambiguity is resolvable (van Deemter, 1998). Preferring one interpretation of an ambiguous statement over another relies on biases (Cairns, 1973). Natural Language Processing (NLP) models have been shown not only to reproduce existing social inequalities (Beukeboom, 2013) but also to amplify them (Dhamala et al., 2021; Tjuatja et al., 2024). Therefore, when faced with an ambiguous input, a well-designed model should remain uncertain to avoid reinforcing biases.

Uncertainty that arises from unresolvable ambiguity can be understood as a form of aleatoric uncertainty (Hora, 1996), since it is inherent to the data and cannot be reduced. Most research considers model uncertainty as a proxy for prediction correctness (Kumar et al., 2019), even though some work has shown that uncertainty can be that model probability distributions encode both aleatoric and epistemic uncertainty (Baan et al., 2024).

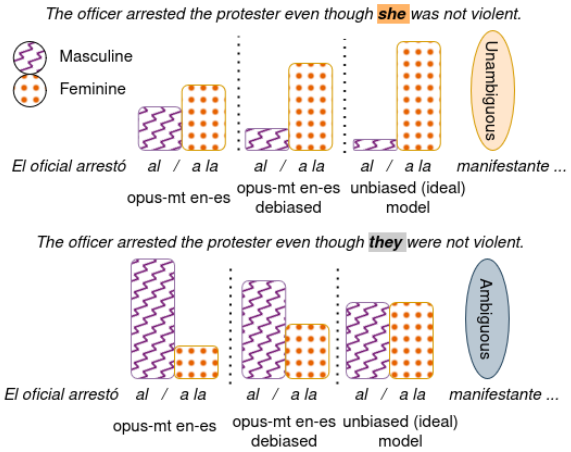

Figure 1: Probabilities for feminine and masculine determiners in a Spanish translation of a sentence containing a noun that is either feminine (referred to as 'she') or ambiguous ('they'), by two existing models and the ideal expected attribution of an unbiased model.

Machine Translation (MT) provides an ideal testing ground for uncertainty caused by ambiguity. That is because languages differ in the number of grammatical categories they explicitly mark, and translating into a richer language requires disambiguation. This allows for explicitly measuring uncertainty caused by ambiguity by changing the specificity of grammatical gender and comparing the resulting translations.

In this work, we leverage uncertainty quantification (UQ) to evaluate bias in cases where the model *should not* be certain about its predictions due to ambiguity in the input. Figure 1 shows two versions of a sentence, with unambiguous (top) or ambiguous (bottom) gender of the noun '*protester*', and the different probabilities assigned to the Spanish translations of the determiner of this noun by

[1]The code will be available at https://anonymous.4open.science/r/uncertainty_bias_ambiguity-8A4C/

an OPUS-MT translation model (Tiedemann and Thottingal, 2020). An ideal model should assign a higher probability for the feminine determiner ('*la*') when the gender is disambiguated by the pronoun, and produce equal probabilities for masculine and feminine translations when the gender is ambiguous. However, in our preliminary analyses, we observed that state-of-the-art MT models, including debiased variants, tend to produce more uniform probability distributions for unambiguous inputs, and less uniform distributions for ambiguous ones. This indicates that model probabilities are influenced by stereotypical associations between protesting and masculinity, causing the models to default to the masculine form even when no gender preference is warranted (bottom), and predict the feminine form with relatively low confidence despite very clear contextual cues (top).

In our analysis we consider semantic UQ metrics for a number of reasons. First, entropy-based uncertainty metrics have been shown to be a suitable measure of ambiguity in human translations (Bangalore et al., 2016), however it has not been applied in MT. Second, quantifying repetition within sampled model outputs, as well as using white-box metrics such as entropy, has been successfully applied to detecting ambiguous inputs in question answering (Cole et al., 2023; Yang et al., 2025). Third, Semantic UQ metrics have been shown to capture linguistic nuance in generation tasks, where the diversity of samples is better estimated if their semantics are taken into account (Cheng and Vlachos, 2024; Farquhar et al., 2024).

Our goal to detect bias using semantic UQ metrics. We focus on extrinsic bias (Goldfarb-Tarrant et al., 2021), which manifests in model outputs that are affected by stereotypes. The currently established metric of gender accuracy evaluates gender bias by measuring how often the model generates the correct gender in the translation (Stanovsky et al., 2019). This metric can only be applied in cases where there exists a *correct gender*, i.e. unambiguous instances. Bias also appears in ambiguous scenarios, therefore we propose to measure it across the board.

To study this systematically, we focus on translating sentences from a language that does not mark gender in nouns and verbs (English) into languages that do (Spanish, French, Ukrainian, Russian, German, Italian and Greek). We use the WINOMT (Stanovsky et al., 2019) as well as the MGENTE (Savoldi et al., 2025) datasets. Both datasets contain instances where the gender is resolvable from context as well as instances where it is not. MGENTE provides alternative translations with gender either explicitly expressed in the target sentence or avoided in the translation. WINOMT includes stereotypical gender roles annotations, and we additionally extend it with manual translations. We compare the ranking of models based on our proposed method with the ranking according to gender accuracy on unambiguous items to prove the usefulness of the Semantic UQ metrics. Then, having a validated metric, we investigate the biases exhibited by the models in ambiguous cases. Our main findings are: 1) the degree of bias correlates with overall model translation accuracy in unambiguous cases, 2) the degree of bias corresponds to translation accuracy at the instance level in ambiguous cases, 3) debiasing effects vary depending on input ambiguity, translation accuracy, and target language.

2 Related Work

Bias in MT has been classified into under-representation and stereotyping, which is persistent even in resolvable settings (Savoldi et al., 2021; Ghosh and Caliskan, 2023). Various methods for debiasing have been proposed (Saunders and Byrne, 2020), however the evaluation of bias is not uniform. Earlier work in gender bias in MT has used intrinsic bias evaluation by testing the embeddings of gendered terms and their association with certain topics (Ramesh et al., 2021), and extrinsic bias evaluation by counting correct translations against a target (Wang et al., 2022). Gender ambiguity in MT has mostly been addressed where it is resolvable (Currey et al., 2022; Robinson et al., 2024), similarly to other types of lexical ambiguity (Barua et al., 2024; Martelli et al., 2025). Existing solutions for unresolvable ambiguous cases (Cho et al., 2019; Gonen and Webster, 2020; Vanmassenhove and Monti, 2021; Farkas and Németh, 2022; Piazzolla et al., 2023) rely on tools or human annotation of gender on specific words, limiting their generalisability to other types of ambiguities. In contrast, some work circumvents the problem of unresolvable gender ambiguity. Lauscher et al. (2023) translate the neutral pronoun while preserving the neutrality in the target language, avoiding exhibiting the bias altogether, while Alhafni et al. (2022) rewrite gender according to the preferences of the user and Vanmassenhove et al. (2018) in-

clude the speaker gender information. Stafanovičs et al. (2020) propose a solution for unresolvable ambiguity by decoupling the task of determining the gender information present in the input from translation, this way circumventing the issue of what should be done when gender information is not present.

When it comes to uncertainty in ML models, methods for distinguishing aleatoric and epistemic uncertainty have been proposed (Hou et al., 2024), however they do not distinguish between data randomness and data ambiguity. Some work has made the link between biases and uncertainty by showing that model uncertainty is linked to the sycophancy bias (Sicilia et al., 2025) as well as the fact that model uncertainty differs when the input mentions different demographic groups (Kuzucu et al., 2025). Ambiguity has also been shown to be reflected in uncertainty in different NLP tasks (Kim, 2025; Cheng and Amiri, 2024), however, these studies construe uncertainty as a signal for ambiguity stemming from low quality data, not including the ambiguity which is an indispensable feature of language.

UQ in MT has been used as a proxy for Quality Estimation (QE). For example, Fomicheva et al. (2020) use MT model uncertainty to estimate translation quality without references, while Glushkova et al. (2021) apply the same technique to the uncertainty of the QE models themselves. Other approaches use UQ to identify difficult instances and enhance training (Zhou et al., 2020; Wei et al., 2020; Wu et al., 2021; Zhan et al., 2023). Wang et al. (2024b) examine zero-shot translation and distinguish between model uncertainty and data uncertainty, however their focus with regard to data uncertainty is on noisy, low-quality training data rather than inherent ambiguity. Ott et al. (2018) show that uncertainty caused by low quality training data is linked to a deterioration of accuracy in beam search with large beams. To the best of our knowledge, no prior UQ-based approach in MT has explored ambiguity as a particular type of data uncertainty.

3 Method

We propose to quantify gender bias in Neural Machine Translation (NMT) models by characterising how gender is assigned to nouns across the predictive distribution. Our bias evaluation method rests on two conditions:

Condition 1: For source sentences with unambiguous gender, an unbiased model should prefer a translation with correct gender inflection compared to an incorrect inflection.

Condition 2: Unbiased models should show higher uncertainty for inputs with ambiguous gender, as compared to unambiguous ones, disregarding the biases in the input.

To do so, we base our approach on recently proposed UQ metrics which are founded on the classic Shannon entropy but take into account similarities between random Monte Carlo samples from the model. Sampling-based metrics are preferred over deterministic decoding methods, as they better capture ambiguity in the ground truth and leverage model stochasticity. We first provide a brief overview of the UQ methods.

Let $\mathcal{Y}$ be a random variable whose value is drawn from the predictive distribution of an NMT model $p(y|x)$. Then entropy is defined as:

$$\mathcal{H}(\mathcal{Y}) = \mathop{\mathbb{E}}_{y\sim\mathcal{Y}}\left[I(y)\right], \tag{1}$$

where I is the *surprisal* of y. In the classic Shannon entropy, $I = -\log p(y)$, but the UQ methods we consider vary in their definition of surprisal.

Semantic Entropy (SE) (Farquhar et al., 2024) is a metric designed to quantify uncertainty by identifying semantic equivalences among output elements. It achieves this by grouping elements y into clusters c of semantically equivalent outputs based on a textual entailment model. To implement this clustering, we use a multilingual mDeberta model (He et al., 2021) finetuned on the Natural Language Inference (NLI) task (Laurer et al., 2022). SE is intended to indirectly capture clusters of translations based on the gender of the subject or object, given that two identical translations that only differ by gender marking would not entail one another.

Given the cluster assignment, the surprisal of an output y under Semantic Entropy is defined as the negative logarithm of the expected probability mass assigned to the cluster c to which y belongs:

$$I_{\text{SE}}(y) = -\log \mathbb{E}_{y'\sim\mathcal{Y}}\left[1[y' \in c]\right], \tag{2}$$

where $\mathcal{Y}$ is the predictive distribution over outputs and $1[\cdot]$ is the indicator function.

Complementing SE, Similarity-sensitive Shannon Entropy (S3E) (Ricotta and Szeidl, 2006; Cheng and Vlachos, 2024) generalizes surprisal by incorporating graded semantic similarity between

outputs rather than binary clusters, capturing subtler translation differences such as gender marking on verbs and adjectives. Specifically, the surprisal of y is the negative logarithm of the expected similarity between y and other outputs y':

$$I_{\text{S3E}}(y) = -\log \mathbb{E}_{y' \sim \mathcal{Y}} \left[\mathcal{S}(y, y') \right], \quad (3)$$

where the similarity function $\mathcal{S} : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$ satisfies $\mathcal{S}(y, y') = 1$ if and only if $y = y'$. Following Cheng and Vlachos (2024), we compute $\mathcal{S}$ as the cosine similarity between sentence embeddings of y and y', generated by a multilingual E5 text embedding model (Wang et al., 2024a).

We also define Gender Entropy (GE), which is calculated like SE but clusters elements based on the gender class of the translated noun. To determine the gender class, we use Spacy[2] and pymorphy2 (Korobov, 2015) morphological parsers. GE provides the most direct way to measure uncertainty with regard to the gender.

To calculate SE, S3E and GE we approximate the model's predictive distribution $p(\cdot \mid x)$ by drawing $N = 128$ samples per input using stochastic decoding. Specifically, we perform ϵ-sampling (Hewitt et al., 2022) with top-p sampling (`epsilon_cutoff = 0.2`). Further details about these UQ methods are given in Appendix B.

We also compute the log probability of each generated sequence by summing the log probabilities of its tokens. Formally, given a sequence $s = (w_1, \ldots, w_T)$ and model output logits scores_t at step t, we calculate

$$\log p(s) = \sum_{t=1}^{T} \log p(w_t \mid w_{<t}, x),$$

where $\log p(w_t \mid w_{<t}, x)$ is obtained by applying the log-softmax to scores_t. This provides a direct likelihood measure of the sequence, used as a simpler alternative to sampling-based surprisal and entropy metrics.

Our gender bias metrics are based on surprisal and entropy from these UQ methods. Thus, to meet **Condition 1**, in an ambiguous setting, an unbiased model should assign lower surprisal to a translation with the correct gender than to an incorrect one. Accordingly, unbiased models should minimize *relative surprisal*, defined as:

$$\Delta I = \frac{I(y_{\text{correct}}) - I(y_{\text{incorrect}})}{\frac{1}{2}\left(I(y_{\text{correct}}) + I(y_{\text{incorrect}})\right)}. \quad (4)$$

[2] https://spacy.io/

To meet the **Condition 2**, unbiased models should show higher uncertainty for an input which is ambiguous with regard to gender, as compared to an input which is unambiguous. It should therefore minimise *relative entropy*, defined as:

$$\Delta \mathcal{H} = \frac{\mathcal{H}(\mathcal{Y}_{\text{unambiguous}}) - \mathcal{H}(\mathcal{Y}_{\text{ambiguous}})}{\frac{1}{2}\left(\mathcal{H}(\mathcal{Y}_{\text{unambiguous}}) + \mathcal{H}\left(\mathcal{Y}_{\text{ambiguous}}\right)\right)}. \quad (5)$$

4 Experimental setup

Data To test our proposed bias metrics, an MT dataset containing information about gender ambiguity and stereotypes in the source sentences is required. We use WINOMT (Stanovsky et al., 2019), which includes annotations on minimal pairs of 1,584 sentences with masculine, feminine or neutral pronouns referring back to stereotypical or anti-stereotypical gender roles. An example of three sentences from the dataset can be seen in Table 1. The gender of the focus noun '*mechanic*' is unambiguous in the first two sentences based on the contextual information (Pronoun M & F), but remains ambiguous in the third on account of the neutral (M/F) pronoun. The Stereotype (column 3) that mechanics are more often men than women either contradicts (row 1) or aligns with (row 2) the disambiguating context (column 2).[3]

Sentence	Pronoun	Stereotype
The **mechanic** called to inform someone that *he* had completed the repair.	M	M
The **mechanic** called to inform someone that *she* had completed the repair.	F	M
The **mechanic** called to inform someone that *they* had completed the repair.	M/F	M

Table 1: WINOMT (Stanovsky et al., 2019) examples.

In addition, we experiment with the contemporaneously proposes dataset MGENTE. The dataset consists of 1,500 sentences, with half of them containing unresolvably ambiguous gender instances, and half where the gender can be determined from the context. While this dataset does not explicitly mark stereotypical gender roles, it provides two alternative translations for each sentence, one with explicit gender marking, and one which avoids specifying the gender in the target language. Table 2 shows examples of both.

The WINOMT dataset has annotations for the focus noun in the input sentence which is ambiguous with regard to gender, whereas in the MGENTE

[3] Please see Appendix A for a broader overview of the biases present in the data.

Sentence	Gender	Translation M/F	Translation G
My second comment is that the previous speakers complained that the European Union pays, but the US is listened to.	M/F	In secondo luogo, le persone che mi hanno preceduto negli interventi hanno lamentato il fatto che l'Unione europea paga, mentre l'America si fa sentire.	In secondo luogo, gli oratori che mi hanno preceduto hanno lamentato il fatto che l'Unione europea paga, mentre l'America si fa sentire.
The former Commissioner, Mr MacSharry, justified this criminal silence by claiming that the beef sector must not be put at risk.	M	MacSharry, in passato già esponente della Commissione, giustificava in tal modo questo silenzio colpevole, sostenendo che non si doveva mettere a repentaglio il settore bovino.	L'ex Commissario MacSharry giustificava in tal modo questo silenzio colpevole, sostenendo che non si doveva mettere a repentaglio il settore bovino.

Table 2: MGENTE (Savoldi et al., 2025) examples.

the target sentences are annotated for gender on the relavant words.

Target Languages We selected target languages that express gender through morphological markers on nouns and adjectives, and sometimes verbs, namely Spanish, French, Ukrainian, Russian, Italian, German and Greek. The target languages vary in their representation within NLP research. A relevant idiosyncrasy of the Russian language is that for some nouns describing professions, even if a feminine version exists, it may be considered derogatory in use (Komova, 2024). For example, 'врач' is the masculine term for '*doctor*', and the alternative feminine 'врачиха' is considered rude, which leads to 'врач' being used even when the doctor is known to be a woman. This results in constructions where masculine nouns are paired with feminine verb forms, or masculine markers are used throughout the sentence. In order to account for this, we also incorporate a list of professions characterized by (lack of) gender marking in WINOMT, based on the categorisation provided by Komova (2024).

Human Translations WINOMT does not contain target translations, thus the accuracy of MT models cannot be directly evaluated. To overcome this, we hired professional translators to translate a set of 100 WINOMT sentences into French, Spanish, Ukrainian and Russian. Each sentence is translated twice, with the focus noun in feminine and masculine variants respectively. They also annotate the translations as Correct or Incorrect with regard to the gender translation in the given context. For instance, when translating the English sentence "*The farmer bought a book from the* ***writer*** *and paid her*" into French, where '*writer*' is the focus noun, the feminine '*l'auteure*' should be marked as Correct, while '*l'auteur*' would be Incorrect. In ambiguous cases, i.e. if the pronoun in the above sentence was '*they*', both gender translations would be Correct. Appendix D provides the translation guidelines and details, and Appendix E discusses the quality of human annotations. We release the translations and correctness annotations to the public to enable further research.

Models We experiment with two commonly used translation models, namely OPUS-MT (Tiedemann and Thottingal, 2020) and M2M100 (Fan et al., 2021), as well as a multilingual LLM model TOWER+ (Rei et al., 2025). OPUS-MT models are NMT models trained on freely available parallel corpora. M2M100 is a many-to-many multilingual translation model, which directly translates between any pair of 100 languages. TOWER+ is an open-weight model which not only exhibits high performance in MT, but also holds strong general capabilities. To examine how effective debiasing is regarding the expectations stated in Section 3, we apply the hard-debiasing method from Iluz et al. (2024) on the OPUS-MT models, which have been shown to lower bias scores on the WINOMT dataset (Stanovsky et al., 2019) while maintaining translation quality. The hard-debiasing method neutralises the biased words in the representation space, so that neutral words are not associated with a specific gender (Bolukbasi et al., 2016). We adopt the most effective debiasing approach from Iluz et al. (2024), which applies debiasing to one-token profession words on the encoder side. Debiasing methods have been shown to have varying effects depending on the evaluation method (Mendelson and Belinkov, 2021; Gonen and Goldberg, 2019), which we also expect to see in our study. See Appendix F for the performance of all models.

5 Research questions

RQ1: Does Semantic Uncertainty Capture Gender Bias? To validate the application of semantic UQ metrics for bias evaluation, we compare their scores with the established gender accuracy metric. Gender accuracy uses the morphological parsers described in Section 3 to determine the focus noun gender in translations. As it relies on gold-standard references, it is applicable only to unambiguous

items and unsuitable for cases with multiple valid gender realisations. We therefore limit this experiment to unambiguous items. We rank all models according to their ΔI scores and compare this ranking to that based on gender accuracy using Kendall's τ and Pearson's r.

RQ2: What does Semantic Entropy Reveal about Bias with Ambiguity? To evaluate model bias in ambiguous settings, which has not yet been explored, we compare models using their $\Delta\mathcal{H}$ scores. To isolate the uncertainty caused by ambiguity from that resulting from poorer model performance, we analyse the relationship between the $\Delta\mathcal{H}$ scores and the translation quality measured by the COMET metric (Rei et al., 2022). Since the WINOMT dataset does not contain gold target translations, we use the 100 professionally annotated items described in Section 4, and the WMT test sets which contain the target languages (Callison-Burch et al., 2012; Bojar et al., 2013, 2014; Koehn et al., 2023; Haddow et al., 2024) for translation quality evaluation.

6 Results

This section presents the results of the bias evaluation using semantic uncertainty metrics.

Lang.	Model	Gender Acc	ΔLog prob	ΔI (S3E)	COMET
ES	OPUS-MT	67.95	0.00	-0.10	84.90
	deb-OPUS-MT	68.13	0.00	-0.13	84.86
	M2M100	70.77	0.00	-0.13	84.53
	TOWER+9B	86.91	0.00	-0.30	89.10
FR	OPUS-MT	64.27	0.01	-0.04	83.56
	deb-OPUS-MT	64.79	0.01	-0.08	83.55
	M2M100	61.66	0.01	-0.07	83.88
	TOWER+9B	75.77	0.02	-0.27	89.76
UK	OPUS-MT	45.34	0.00	-0.03	70.79
	deb-OPUS-MT	46.12	0.00	-0.03	70.79
	M2M100	47.76	0.00	-0.02	76.52
	TOWER+9B	57.58	0.01	-0.12	92.26
RU	OPUS-MT	48.57	0.00	0.00	79.37
	deb-OPUS-MT	48.42	0.00	-0.03	79.36
	M2M100	48.49	0.00	-0.03	80.49
	TOWER+9B	51.44	0.02	-0.09	90.92

Table 3: Gender Accuracy, Δ Log Probability, and ΔI (S3E) on Unambiguous instances of WINOMT, COMET scores on WMT test sets (see Appendix F for details).

RQ1: Model Rankings According to Semantic Surprisal and Gender Accuracy Correlate. The results of the first experiment on WINOMT are presented in Table 3. We find that while ΔLog prob does not correlate with the gender accuracy ranking, ΔI (S3E) shows a statistically significant negative correlation with gender accuracy (Spearman's $\rho = -0.84$; see Appendix M). We attribute the effectiveness of this metric in distinguishing Correct from Incorrect gender translations to its ability to capture gender information beyond noun morphology, including verb inflections and agreement, through its embedding representations. The flexibility of S3E enables it to encode nuances that GE does not. For example, when translating a sentence with a feminine pronoun into Russian, OPUS-MT generates sentences with a masculine noun and a verb that is either feminine or masculine (e.g., "Перевозчик поблагодарила (fem) / поблагодарил (masc)" "*The courier thanked*"). This is reflected in a higher $\mathcal{H}$ score assigned to this sentence by S3E: $\mathcal{H} = 0.65$, than GE: $\mathcal{H} = 0.00$, as the variation in verb inflections is only captured by S3E.

The very strong negative correlation between ΔI (S3E) and gender accuracy thus validates the core component of our proposed metric for evaluating bias in machine translation. Our findings indicate that we can rank models according to their bias in unambiguous settings with our proposed metric as well as it can be done with gender accuracy, but without the need for explicit gender annotations.

We do not find a correlation between the surprisal scores and gender accuracy in the MGENTE dataset (please see Appendix K for the scores). There are a number of factors that explain this. First, in WINOMT the Incorrect translations are explicitly the wrong gender, while in MGENTE the Incorrect translations contain neutral expressions which do not match the explicit gender markers in the remainder of the same sentence (please see Translation N in the top example in Table 2). This is not strictly grammatically incorrect, only pragmatically less preferred, which is expected to elicit less surprisal from the model. Second, the MGENTE dataset is not designed to include bias via stereotypes, as the gendered content of the sentence neither strongly aligns with nor contradicts the gender marking. This leads to smaller differences between the predicted likelihood of different gender expressions, with the highest absolute value of ΔH reaching 0.25, compared to 0.74 on WINOMT.

RQ2: Translation Accuracy Affects the Bias–Entropy Relationship Differently Across Levels of Analysis. Having validated the S3E metric in the first experiment, we investigate the results of $\Delta\mathcal{H}$ (S3E) as a bias metric. Table 4 illus-

trates that all models exhibit the desired negative $\Delta\mathcal{H}$ for some languages in the WINOMT dataset. TOWER+9B, which performs best in translation accuracy, also exhibits the desired unbiased behavior with consistently negative $\Delta\mathcal{H}$, suggesting that better models are less biased. Surprisingly, when it comes to ambiguous instances, contrary to the unambiguous cases, the languages where the models perform better on translation accuracy, namely Spanish and French, are not generally less gender biased under ambiguity (model rankings according to all metrics are available in Appendix L). Nonetheless, we observe the expected effect of debiasing on the Spanish and French translations (lower $\Delta\mathcal{H}$ for deb-OPUS-MT).

Lang.	Model	Unamb	Amb	$\Delta\mathcal{H}$
ES	OPUS-MT	1.23	1.12	0.09
	deb-OPUS-MT	0.97	0.89	0.08
	M2M100	1.79	1.45	0.19
	TOWER+9B	1.67	1.82	**-0.15**
FR	OPUS-MT	1.79	1.43	0.20
	deb-OPUS-MT	1.21	1.08	0.11
	M2M100	3.22	2.78	0.14
	TOWER+9B	2.72	3.46	**-0.74**
UK	OPUS-MT	1.96	2.16	**-0.10**
	deb-OPUS-MT	1.98	2.15	**-0.09**
	M2M100	2.05	2.28	**-0.11**
	TOWER+9B	4.12	4.85	**-0.73**
RU	OPUS-MT	1.56	1.68	-0.08
	deb-OPUS-MT	1.05	0.97	0.08
	M2M100	1.83	2.29	**-0.25**
	TOWER+9B	3.15	3.63	**-0.48**

Table 4: Unambiguous and Ambiguous $\mathcal{H}$ (S3E) on WINOMT. Statistically significant correlations between $\Delta\mathcal{H}$ and COMET scores on the instance level are in **bold**.

The results on the MGENTE dataset corroborate the finding that the best performing model (TOWER+9B) exhibits the least amount of bias (see Table 17 in Appendix K). The effects are smaller compared to the WINOMT dataset, and debiasing does not show an effect on MGENTE, which may be explained by the fact that the dataset is not explicitly stereotypical and less bias can be observed.

In Figure 2, results on WINOMT are grouped by COMET score bins for a more fine-grained analysis at the instance level. For the models with negative $\Delta\mathcal{H}$ (S3E) scores, $\Delta\mathcal{H}$ is typically most pronounced for the highest-accuracy translations (e.g. ambiguous scores for TOWER+9B-UK in Bin 3 are substantially higher than B1 and B2). Although debiasing does not reduce the overall $\Delta\mathcal{H}$ score for Ukrainian (see Table 4), it results in the largest improvement in the highest-quality translations: in bin B3, ambiguous $\mathcal{H}$ scores for deb-OPUS-MT-UK are notably higher than those of the original model. This improvement is further supported by a substantial 8.41% drop in masculine focus noun inflections for Ukrainian, compared to 0.88–2.49% for other languages. We hypothesise that this is due to the limited training data in Ukrainian, which may lead to a less stable model that performs worse overall but is more responsive to debiasing in high-quality outputs. The relationship between translation accuracy and bias under ambiguity appears to differ depending on the level of analysis: *between* models, higher accuracy does *not* imply lower bias on ambiguous instances, whereas *within* models, higher-accuracy instances tend to show *lower* bias.

Qualitative Analysis A qualitative analysis of the example in Table 1, presented with corresponding $\mathcal{H}$ values in Table 5, corroborates the quantitative findings in Figure 2. When translating the sentence in row 1, across all four target languages used in WINOMT, OPUS-MT models consistently produce only the masculine variants of the focus noun ('*El mecánico*', '*Le mécanicien*', 'Механик' and 'Механік'). In the anti-stereotypical case (row 2), all languages except Russian include both masculine and feminine forms ('*La mecánica*', '*La mécanicienne*' and 'Механіка'), indicating that these models are sensitive to the masculine stereotype even when the referent in the context is clearly feminine. For Russian, the models fail to generate any feminine constructions, even when the context is unambiguously feminine. This difference is evident in the $\mathcal{H}$ (S3E) scores, which are higher in row 2 than row 1 for the first three languages. Moreover, in the ambiguous case (row 3), all OPUS-MT models produce only masculine nouns, regardless of language. Consequently, $\mathcal{H}$ (S3E) scores are generally higher for the unambiguous cases (mean of rows 1 and 2) than for the ambiguous case (row 3), except for Russian, where $\mathcal{H}$ remains low across all conditions. These observations are consistent with expectations for biased models: they default to stereotypical gender realisations when the pronoun is ambiguous and sometimes even when the context clearly suggests an anti-stereotypical interpretation.

The qualitative analysis also reveals interesting effects of debiasing. In the anti-stereotypical case in row 2, debiasing increases the number of feminine constructions generated in Spanish (from

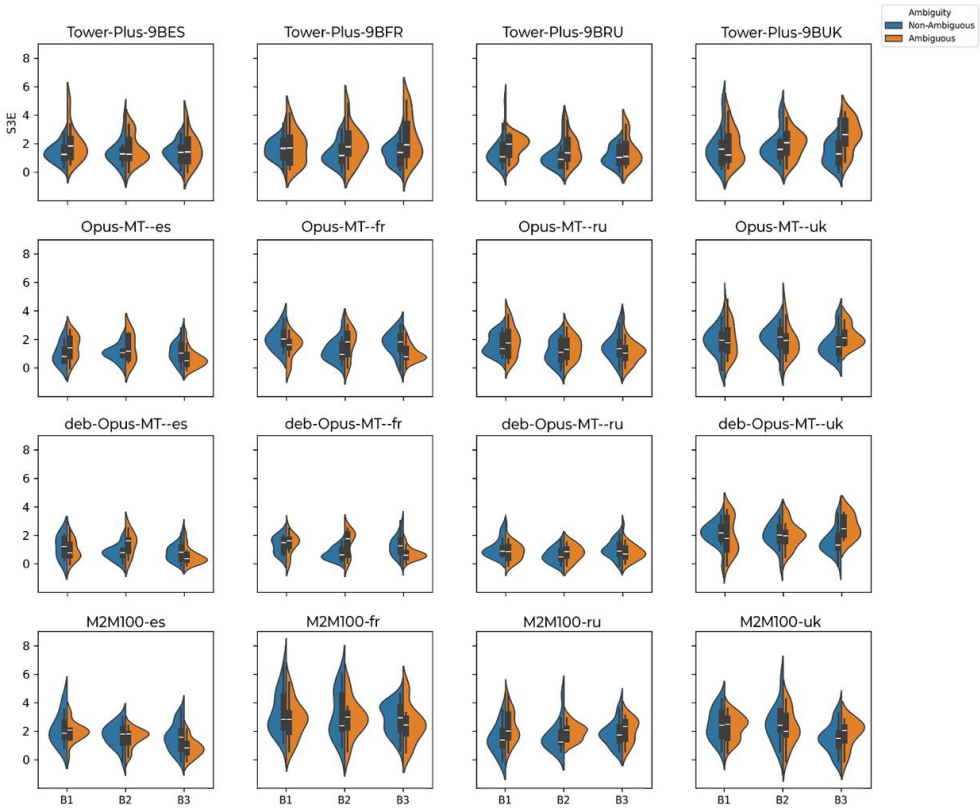

Figure 2: Violin plots of $\mathcal{H}$ (S3E) on ambiguous and unambiguous WINOMT instances. Binned by low (B1), medium (B2) and high (B3) COMET scores against human translations.

Sentence	**ES** → deb		**FR** → deb		**UK** → deb		**RU** → deb	
The **mechanic** called to inform someone that *he* had completed the repair.	0.75	0.82	0.00	0.00	2.41	1.57	0.33	0.31
The **mechanic** called to inform someone that *she* had completed the repair.	1.64	1.85	0.53	0.56	2.75	2.05	0.00	0.00
The **mechanic** called to inform someone that *they* had completed the repair.	0.74	0.87	0.00	0.00	2.38	2.03	0.37	0.37

Table 5: WINOMT examples with $\mathcal{H}$ (S3E) values, for OPUS-MT (left) and deb-OPUS-MT (right) models.

43/128 to 55/128, corresponding to a slight increase in $\mathcal{H}$, as feminine forms remain a minority) and Ukrainian (from 72/128 to 128/128, reflected in a decrease in $\mathcal{H}$). No notable changes are observed for French or Russian, consistent with stable $\mathcal{H}$ (S3E) scores. When it comes to the ambiguous pronoun (row 3), the debiased models continue to generate only masculine variants of '*mechanic*' in Spanish, French and Russian, with $\mathcal{H}$ remaining largely unchanged. In contrast, all debiased model outputs in Ukrainian include a feminine translation of the noun, corresponding to a decrease in $\mathcal{H}$ from OPUS-MT to debiased OPUS-MT in row 3. This pattern illustrates that when debiasing leads to overgeneration of feminine morphology in ambiguous contexts, our proposed metric flags this as increased bias (positive $\Delta\mathcal{H}$), indicating that such changes are not deemed as improvements.

7 Conclusion

In this work, we apply distribution-level UQ to evaluate bias in MT models. This method complements gender accuracy, particularly where gender accuracy is inapplicable. Specifically, it captures the more subtle manifestations of gender bias that arise when models show a preference for one gender in ambiguous contexts. Our overall contribution is the novel use of UQ as a bias metric in MT, which 1) does not rely on gender references, 2) is general and captures multiple types of bias, 3) is validated by the established metric of gender accuracy, and 4) provides new insights into biased behavior in ambiguous contexts, a setting not previously studied. Future work will extend the proposed bias evaluation method to tasks beyond translation.

Limitations

This study is limited to Romance and Slavic languages, not including many other language families which mark gender and express stereotypes in diverse ways. While we tried to account for language differences by including different names for different target languages, accounting for specific masculine-only nouns in Russian, debiasing with language-specific vocabularies, etc., some linguistic idiosyncracies are still not accounted for, such as the fact that profession stereotypes are defined in English and may apply differently in different regions. Finally, our work is limited to two grammatical genders, and treats 'they' as a neutral pronoun that may refer to any gender, however we do not study the interpretation of the pronoun as referring to non-binary people specifically. Further directions include applying UQ to ambiguity detection, which could enable more gender-inclusive translations through morphological doubling, where both masculine and feminine morphemes are included for gender neutrality. Future work should address these directions.

Ethics Statement

The models used in this study, like all ML models, can be biased and make mistakes, including in gender attribution. Our contribution aims to specifically tackle masculine and feminine gender stereotypes via more stringent evaluation metrics, in order to avoid the perpetuation of gender bias.

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. User-centric gender rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting predictive probabilities: Model confidence or human label variation? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian's, Malta. Association for Computational Linguistics.

Srinivas Bangalore, Bergljot Behrens, Michael Carl, Maheshwar Ghankot, Arndt Heilmann, Jean Nitzke, Moritz Schaeffer, and Annegret Sturm. 2016. *Syntactic Variance and Priming Effects in Translation*, pages 211–238. Springer International Publishing, Cham.

Josh Barua, Sanjay Subramanian, Kayo Yin, and Alane Suhr. 2024. Using language models to disambiguate lexical choices in translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4837–4848, Miami, Florida, USA. Association for Computational Linguistics.

Camiel J Beukeboom. 2013. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social cognition and communication*, pages 313–330.

Ondrej Bojar, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia, editors. 2013. *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors. 2014. *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Helen S Cairns. 1973. Effects of bias on processing and reprocessing of lexically ambiguous sentences. *Journal of Experimental Psychology*, 97(3):337.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia, editors. 2012. *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Montréal, Canada.

Jiali Cheng and Hadi Amiri. 2024. FairFlow: Mitigating dataset biases through undecided learning for natural language understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21960–21975, Miami, Florida, USA. Association for Computational Linguistics.

Julius Cheng and Andreas Vlachos. 2024. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2115–2128, St. Julian's, Malta. Association for Computational Linguistics.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the*

First Workshop on Gender Bias in Natural Language Processing, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Anna Farkas and Renáta Németh. 2022. How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points. *Social Sciences & Humanities Open*, 5(1):100239.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Alan Garnham, Svenja Vorthmann, and Karolina Kaplanova. 2021. Implicit consequentiality bias in english: A corpus of 300+ verbs. *Behavior Research Methods*, 53:1530–1550.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with english llms: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 901–912.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.

Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors. 2024. *Proceedings of the Ninth Conference on Machine Translation*. Association for Computational Linguistics, Miami, Florida, USA.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephen C. Hora. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2):217–223. Treatment of Aleatory and Epistemic Uncertainty.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Bar Iluz, Yanai Elazar, Asaf Yehudai, and Gabriel Stanovsky. 2024. Applying intrinsic debiasing on downstream tasks: Challenges and considerations for machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14914–14921, Miami, Florida, USA. Association for Computational Linguistics.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised Mary because _he_? implicit causality bias and its interaction with explicit cues in LMs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4859–4871, Online. Association for Computational Linguistics.

Hazel H. Kim. 2025. How ambiguous are the rationales for natural language reasoning? a simple approach to handling rationale uncertainty. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10047–10053, Abu Dhabi, UAE. Association for Computational Linguistics.

Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors. 2023. *Proceedings of the Eighth Conference on Machine Translation*. Association for Computational Linguistics, Singapore.

Liliana Komova. 2024. *Gender, Language and Perception: Linguistic Inclusivity in Russian*. Ph.D. thesis, Università Ca'Foscari Venezia.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *Analysis of Images, Social Networks and Texts*, pages 320–332, Cham. Springer International Publishing.

Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Selim Kuzucu, Jiaee Cheong, Hatice Gunes, and Sinan Kalkan. 2025. Uncertainty as a fairness measure. *Journal of Artificial Intelligence Research*, 81.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? how commercial machine translation fails to handle (neo-)pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Federico Martelli, Stefano Perrella, Niccolò Campolungo, Tina Munda, Svetla Koeva, Carole Tiberius, and Roberto Navigli. 2025. Dibimt: A gold evaluation benchmark for studying lexical ambiguity in machine translation. *Computational Linguistics*, pages 1–71.

Michael Mendelson and Yonatan Belinkov. 2021. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1557, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mante S. Nieuwland and Jos J.A. Van Berkum. 2006. Individual differences and contextual bias in pronoun resolution: Evidence from erps. *Brain Research*, 1118(1):155–167.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3953–3962. PMLR.

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2023. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *HERMES*, (63):209–225.

Krithika Ramesh, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in Hindi-English machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23, Online. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André FT Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *arXiv preprint arXiv:2506.17080*.

Carlo Ricotta and Laszlo Szeidl. 2006. Towards a unifying approach to diversity measures: Bridging the gap between the shannon entropy and rao's quadratic index. *Theoretical Population Biology*, 70(3):237–243.

Kevin Robinson, Sneha Kudugunta, Romina Stella, Sunipa Dev, and Jasmijn Bastings. 2024. MiTTenS:

A dataset for evaluating gender mistranslation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4115–4124, Miami, Florida, USA. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Danielle Saunders and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 85–93, Tampere, Finland. European Association for Machine Translation.

Beatrice Savoldi, Eleonora Cupin, Manjinder Thind, Anne Lauscher, Andrea Piergentili, Matteo Negri, and Luisa Bentivogli. 2025. mgente: A multilingual resource for gender-neutral language and translation. *arXiv preprint arXiv:2501.09409*.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Anthony Sicilia, Mert Inan, and Malihe Alikhani. 2025. Accounting for sycophancy in language model uncertainty estimation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7851–7866, Albuquerque, New Mexico. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, 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 Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Kees van Deemter. 1998. Ambiguity and idiosyncratic interpretation. *Journal of Semantics*, 15(1):5–36.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Eva Vanmassenhove and Johanna Monti. 2021. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online. Association for Computational Linguistics.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wenxuan Wang, Wenxiang Jiao, Shuo Wang, Zhaopeng Tu, and Michael R. Lyu. 2024b. Understanding and mitigating the uncertainty in zero-shot translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:4894–4904.

Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Luxi Xing, and Weihua Luo. 2020. Uncertainty-aware semantic augmentation for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2724–2735, Online. Association for Computational Linguistics.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. 2021. Uncertainty-aware balancing for multilingual and multi-domain neural machine translation training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7291–7305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5846–5863, Albuquerque, New Mexico. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, Cuilian Zhang, Lidia S. Chao, and Min Zhang. 2023. Test-time adaptation for machine translation evaluation by uncertainty minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

807–820, Toronto, Canada. Association for Computational Linguistics.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944, Online. Association for Computational Linguistics.

A Effects of Different Biases

This dataset also contains other linguistic phenomena that were not explicitly annotated in its released version. Thus, we automatically annotate additional linguistic bias cues, namely subject, recency, implicit causality, and person names, using syntactic parses with Spacy(`https://spacy.io/`). We release the additional annotations to the public for reproducibility.

For the Subject bias, following the literature on human biases for coherence (Nieuwland and Van Berkum, 2006), we hypothesise that models may assume coreference between the subject of the main clause (often the focus noun) and the subject of the complement clause (often the pronoun). In the example in Table 6 '*the mechanic*' is the subject, therefore the Subject bias primes an interpretation in which the gender of the subject aligns with that of the subsequent pronoun (M in row 1, F in row 2). Furthermore, person names have been shown to have a strong effect on pronoun resolution (Saunders and Olsen, 2023). To assess the impact of person names on gender translation, we augment the dataset with common feminine and masculine names matching the gender of the pronoun, selected for their cross-linguistic familiarity (see Appendix C). As an example, when translating into French, the second sentence from Table 6 would read "*The mechanic Anne called to inform someone that she had completed the repair.*"

Recency bias elicits the attribution of the gender of the most recent noun phrase to the following pronoun (Gautam et al., 2024). However, in our example, the most recent noun phrase is '*someone*', so the focus noun is unaffected. The implicit causality (IC) bias induces an expectation in humans and LMs that when an IC verb is followed by a causal connective, the following pronoun will corefer with either the Subject or the Object of the IC verb, depending on the causal inference (Kementchedjhieva et al., 2021). For instance, a clause with the verb '*call*', when followed by an explanation starting with '*because*', is expected to be followed by a mention of the *caller* rather than the *callee*. We use an IC verb corpus (Garnham et al., 2021) for the annotation and find that about 10% of the dataset contains IC structures. The sentences in Table 6 do not include a causal complement, hence the IC bias does not apply.

To evaluate model bias, we assess how bias cues influence the diversity of gender markers in translations. Specifically, we perform an analysis of variance (ANOVA) to examine the effect of bias cues (independent variables) on normalised entropy measures norm-$\mathcal{H}$ (S3E), norm-$\mathcal{H}$ (SE), and norm-$\mathcal{H}$ (GE) (dependent variables), using T-tests for significance. This analysis incorporates linguistic biases not previously explored in WINOMT.

Our expectation is that the entropy of unbiased models should not be affected by the presence of bias cues. Thus, we define *normalised entropy*, which compares the $\mathcal{H}$ of a source sentence x to the average entropy across its *contrast set* $\mathcal{G}_x$. $\mathcal{G}_x$ is a group of minimally different sentences that are identical to x apart from the pronoun (e.g., '*she*', '*he*', '*they*'), including x itself. The three sentences in Table 1 comprise a $\mathcal{G}_x$. Formally:

$$\text{norm-}\mathcal{H}(x) = \frac{\mathcal{H}(\mathcal{Y}_x)}{\frac{1}{|\mathcal{G}_x|}\sum_{x' \in \mathcal{G}_x} \mathcal{H}(\mathcal{Y}_{x'})}, \quad (6)$$

This formulation isolates variation in $\mathcal{H}(\mathcal{Y}_x)$ attributable specifically to gender by holding all other lexical, syntactic, and semantic content constant across the contrast set.

The results of the experiment presented in Table 7 show that most bias cues in the data have a significant effect on the variance of norm-$\mathcal{H}$ (S3E), indicating that the tested models exhibit various social and linguistic biases.[4] The results corroborate

[4] norm-$\mathcal{H}$ (S3E) shows the strongest sensitivity to bias cues compared to norm-$\mathcal{H}$ (SE) and norm-$\mathcal{H}$ (GE). We also experiment with unnormalised $\mathcal{H}$ scores, the results of which

Sentence	Pronoun	Stereotype	Subject	Recency	IC
The **mechanic** called to inform someone that *he* had completed the repair.	M	M	M	N	N
The **mechanic** called to inform someone that *she* had completed the repair.	F	M	F	N	N
The **mechanic** called to inform someone that *they* had completed the repair.	N	M	N	N	N

Table 6: WINOMT (Stanovsky et al., 2019) examples with additional annotations of bias cues.

Lang.	Model	Names	Recency		Implicit Causality				Stereotype				Subject		Pronoun				Default M		Ambiguity
			F	M	S F	S M	O F	O M	S F	S M	O F	O M	F	M	S F	S M	O F	O M	S	O	
es	OPUS-MT	**0.41**	**0.41**	-0.05	**0.25**	**-0.19**	**0.24**	**-0.33**	0.06	0.10	**0.13**	**0.17**	**0.38**	**-0.21**	**0.24**	**-0.31**	**0.29**	**-0.13**	N/A	N/A	**-0.18**
	deb-OPUS-MT	**-0.05**	**0.30**	**-0.11**	**0.27**	**-0.20**	**0.16**	**-0.38**	0.05	**0.14**	**0.08**	0.05	**0.49**	**-0.14**	**0.39**	**-0.24**	**0.14**	**-0.21**	N/A	N/A	**-0.10**
	M2M100	**0.14**	**0.33**	**-0.11**	**0.29**	**-0.42**	**0.28**	**-0.25**	**0.18**	**0.27**	**0.12**	0.07	**0.51**	-0.04	**0.52**	**-0.08**	0.04	**-0.35**	N/A	N/A	**-0.11**
fr	OPUS-MT	**0.54**	**0.42**	**0.16**	0.05	**-0.34**	0.06	**-0.24**	**0.43**	**0.45**	**0.26**	**0.21**	**0.16**	**-0.14**	**0.16**	**-0.17**	**0.20**	-0.04	N/A	N/A	**-0.29**
	deb-OPUS-MT	**0.19**	**0.22**	0.02	0.05	**-0.25**	**0.17**	**-0.11**	**0.23**	**0.32**	**0.23**	**0.13**	**0.24**	-0.03	**0.31**	**-0.08**	0.01	**-0.14**	N/A	N/A	**-0.12**
	M2M100	**-0.12**	0.11	**-0.23**	**0.50**	0.09	**0.49**	0.03	-0.03	0.11	**0.21**	0.08	**0.70**	**0.31**	**0.47**	0.05	**-0.08**	**-0.40**	N/A	N/A	0.06
uk	OPUS-MT	**-0.27**	0.00	**-0.11**	**0.26**	-0.02	**0.19**	**0.13**	-0.11	**0.17**	0.03	**-0.14**	**0.27**	**0.21**	**0.22**	**0.13**	**-0.11**	**-0.23**	N/A	N/A	0.06
	deb-OPUS-MT	**-0.43**	-0.06	**-0.12**	**0.41**	0.00	**0.18**	0.07	**-0.24**	0.01	0.01	**-0.17**	**0.23**	**0.17**	**0.16**	**0.10**	**-0.11**	**-0.17**	N/A	N/A	**0.09**
	M2M100	**0.08**	-0.04	**-0.18**	**0.27**	0.02	**0.33**	**0.18**	**0.19**	**0.37**	-0.03	**-0.17**	**0.53**	**0.34**	**0.50**	**0.26**	**-0.30**	**-0.42**	N/A	N/A	**0.11**
ru	OPUS-MT	0.01	**-0.28**	**-0.41**	0.00	-0.12	0.08	**-0.10**	**-0.19**	**-0.20**	**-0.41**	**-0.39**	**0.07**	-0.03	**0.14**	0.03	**-0.10**	**-0.24**	**-0.40**	0.04	**0.35**
	deb-OPUS-MT	**0.23**	**-0.10**	**-0.17**	0.05	0.03	0.05	-0.03	**-0.10**	-0.04	**-0.11**	**-0.14**	**0.09**	**0.04**	**-0.13**	0.03	**-0.08**	**-0.13**	**-0.26**	0.00	**0.13**
	M2M100	**-0.74**	**-0.16**	**-0.21**	0.10	-0.04	**-0.12**	**-0.25**	**0.12**	**0.12**	**-0.20**	**-0.18**	**0.06**	-0.02	**0.08**	0.01	**-0.07**	**-0.11**	**-0.20**	**-0.22**	**0.18**

Table 7: ANOVA results: single effects of bias cues (**F**eminine, **M**asculine, **S**ubject and **O**bject) on norm-$\mathcal{H}$ (S3E). Values correspond to effect coefficients (deviations from a reference group). **Boldface** indicates statistical significance ($p < 0.05$). The sign of the values indicates whether the presence of the variable increases (positive) or decreases (negative) the mean $\mathcal{H}$ of the group containing the given variable value. Reference group is N for all columns except: 'no name' for Names, 'no default' for Default M, 'unambiguous' for Ambiguity.

previous findings. The high absolute coefficient values in the Names column indicate that person names have an effect on gender translation even when a disambiguating pronoun is present. This is in line with Saunders and Olsen (2023), who have shown that both pronouns and names induce gender bias and are often not sufficient for full disambiguation. Secondly, the fact that some Russian nouns have a default masculine grammatical gender regardless of the context (Komova, 2024) is reflected in significant decrease in gender diversity for sentences containing such nouns (negative coefficients in the Default M columns indicating lower norm-$\mathcal{H}$). Thirdly, we observe that masculine biases generally reduce norm-$\mathcal{H}$ (negative coefficients in the M columns), while feminine biases tend to increase it (positive coefficients in the F columns). This suggests a general default toward masculine translations in models, with outputs becoming more similar under masculine biases and more varied under feminine ones. This finding aligns with Kuzucu et al. (2025), who show that model uncertainty is typically higher for minority groups. The results in the Ambiguity column in Table 7 indicate that norm-$\mathcal{H}$ increases for Ukrainian and Russian (positive coefficients), but not for Spanish and French, which mirrors the results from Section 6. Higher norm-$\mathcal{H}$ for Ambiguous items (or negative $\Delta\mathcal{H}$) is expected for an unbiased model.

Table 7 further demonstrates the impact of debiasing, with effect sizes typically reduced (smaller absolute coefficients) in debiased models versus non-debiased ones across languages, supporting the partial effectiveness of debiasing.

are less comparable across metrics, bias types, and models. The full results are presented in Appendix G. All trends observed for norm-$\mathcal{H}$ (S3E) are also present in norm-$\mathcal{H}$ (SE) and norm-$\mathcal{H}$ (GE), as well as for unnormalised $\mathcal{H}$.

B Further UQ discussion

Farquhar et al. (2024) does not define a per-element surprisal; the original definition computes Shannon entropy over clusters. $\mathcal{Y}$ are mapped to clusters $\mathcal{C}$, and SE is:

$$\mathcal{H}_{\text{SE}}(\mathcal{C}) = -\mathop{\mathbb{E}}_{c\sim\mathcal{C}} \log p(c|x).$$

Surprisal is thus defined for a cluster instead of an element, but it is easy to show that our per-element surprisal obtains equivalent entropy as the original definition.

Cheng and Vlachos (2024) introduce a hyperparameter α which is applied as an exponent to the similarity function. This is used to rescale $\mathcal{S}$ for more favorable performance on benchmarks. We tune α for the highest correlation between S3E and the entropy of the gender labels assigned to the nouns in question by the morphological parser. This way we aim for $\mathcal{H}$ (S3E) to capture gender variation, with higher values of $\mathcal{H}$ (S3E) indicating more diversity in the gender morphemes.

We also experiment with alternative similarity metrics to cosine similarity, such as Euclidean Distance, Chebyshev Distance, Manhattan Distance, Minkowski Distance, etc. but find no notable differences and use the simplest option of Cosine Similarity in the paper.

C Annotation of Names

Table 8 presents the names used for expanding the WINOMT dataset to include commonly used person names for masculine and feminine genders in French, Spanish, Ukrainian and Russian.

	Spanish	French	Ukrainian	Russian
F	Carla, Lucía, María, Sofía, Ana	Anne, Camille, Élise, Sophie, Claire	Anna, Olena, Sofiia, Natalka, Oksana	Anna, Maria, Elena, Svetlana, Natalia
M	Gabriel, Carlos, Luis, Javier, Miguel	Victor, Julien, Antoine, Lucas, Mathieu	Ivan, Andrii, Taras, Oleh, Mykola	Ivan, Dmitry, Alexei, Sergey, Nikolai

Table 8: The names used for extending the WINOMT dataset for different target languages.

D Translation Guidelines and Translator Recruitment Details

Professional translators were hired and paid £0.2 per word as requested. The translation guidelines are as follows:

Please translate the following sentences into [French/Spanish/Ukrainian/Russian]. Please provide two translations for each sentence, with different gender inflections for the highlighted noun —one masculine and one feminine. Please keep the pronoun true to the original to the extent that it is possible. Please use the translations that appear natural to you as a native speaker rather than using forms that are prescribed[5]*. Please note whether both translations are acceptable in terms of the gender of the noun matching the contextual information, and if they are not, which one is the correct one:*

- The **developer** argued with the designer because she did not like the design.
- The undergraduate presented the **administrator** with a petition that they had organized.

E Human Annotation Quality

The annotations of Correct and Incorrect sentences by professional translators are compared to the gold standard gender annotations in the dataset, and across all four translators range between 93.17 and 93.27 in Cohen's κ scores, confirming that apart from some linguistic idiosyncrasies of each language (e.g. 'victim' in Spanish is always feminine and so regardless of the contextualising pronoun will take the same form), the annotators agree on which sentences should be correctly translated in which gender.

F Overall Model Performance

Table 9 presents the performance of the models used in this study in terms of the COMET metric[6] (Rei et al., 2022) on WMT datasets which contain the target languages (Callison-Burch et al., 2012; Bojar et al., 2013, 2014; Koehn et al., 2023; Haddow et al., 2024). The models are run on a single NVIDIA TU102 GPU.

[5] This instruction was added due to inquiries by translators about the use of formal or dominant dialects as well as language policy changes influencing how they might translate some words.

[6] https://huggingface.co/Unbabel/wmt22-comet-da

Dataset	Model	es	fr	uk	ru	de	it	el
newstest2012	Opus-MT	84.52	82.21	–	–	–	–	–
	deb-Opus-MT	84.47	82.22	–	–	–	–	–
	m2m100	84.23	82.69	–	–	–	–	–
	Tower+9B	86.61	84.86	–	–	–	–	–
newstest2013	Opus-MT	85.28	83.45	–	–	–	–	–
	deb-Opus-MT	85.24	83.44	–	–	–	–	–
	m2m100	84.83	83.66	–	–	–	–	–
	Tower+9B	87.14	86.27	–	–	–	–	–
newstest2014	Opus-MT	–	85.01	–	87.44	–	–	–
	deb-Opus-MT	–	85.00	–	87.44	–	–	–
	m2m100	–	85.28	–	71.95	–	–	–
	Tower+9B	–	88.71	–	93.04	–	–	–
wmttest2023	Opus-MT	–	–	74.58	79.02	–	–	–
	deb-Opus-MT	–	–	74.58	79.02	–	–	–
	m2m100	–	–	73.58	80.23	–	–	–
	Tower+9B	–	–	86.24	83.75	–	–	–
wmttest2024	Opus-MT	–	–	66.99	71.64	–	–	–
	deb-Opus-MT	–	–	67.00	71.63	–	–	–
	m2m100	–	–	79.45	89.30	–	–	–
	Tower+9B	–	–	87.07	85.69	–	–	–
newstest2009	Opus-MT	–	–	–	–	81.30	84.65	–
	deb-Opus-MT	–	–	–	–	81.24	84.58	–
	m2m100	–	–	–	–	82.29	85.04	–
	Tower+9B	–	–	–	–	85.33	87.78	–
newstest2024	Opus-MT	–	–	–	–	–	–	78.41
	deb-Opus-MT	–	–	–	–	–	–	77.90
	m2m100	–	–	–	–	–	–	78.58
	Tower+9B	–	–	–	–	–	–	66.23
mean	Opus-MT	84.90	83.56	70.79	79.37	81.30	84.65	78.41
	deb-Opus-MT	84.86	83.55	70.79	79.36	81.24	84.58	77.90
	m2m100	84.53	83.88	76.52	80.49	82.29	85.04	78.58
	Tower+9B	86.88	86.61	86.66	87.49	85.33	87.78	66.23

Table 9: COMET scores on WMT test sets for the models used.

G ANOVA Results

Table 10 presents the ANOVA results for S3E, SE and GE metrics. Table 11 presents the ANOVA results without normalising the $\mathcal{H}$ values.

Lang.	Model	Names	Recency		Implicit Causality				Stereotype				Subject		Context				Default M		Ambiguity
			F	M	S F	S M	O F	O M	S F	S M	O F	O M	F	M	S F	S M	O F	O M	S	O	
										S3E											
	OPUS-MT	**0.41**	**0.41**	-0.05	**0.25**	**-0.19**	**0.24**	**-0.33**	0.06	0.10	**0.13**	**0.17**	**0.38**	**-0.21**	**0.24**	**-0.31**	**0.29**	**-0.13**	N/A	N/A	**-0.18**
es	deb-OPUS-MT	**-0.05**	**0.30**	**-0.11**	**0.27**	**-0.20**	**0.16**	**-0.38**	0.05	**0.14**	**0.08**	0.05	**0.49**	**-0.14**	**0.39**	**-0.24**	**0.14**	**-0.21**	N/A	N/A	**-0.10**
	M2M100	**0.14**	**0.33**	**-0.11**	**0.29**	**-0.42**	**0.28**	**-0.25**	**0.18**	**0.27**	**0.12**	0.07	**0.51**	-0.04	**0.52**	**-0.08**	0.04	**-0.35**	N/A	N/A	**-0.11**
	OPUS-MT	**0.54**	**0.42**	**0.16**	0.05	**-0.34**	0.06	**-0.24**	**0.43**	**0.45**	**0.26**	**0.21**	**0.16**	**-0.14**	**0.16**	**-0.17**	**0.20**	-0.04	N/A	N/A	**-0.29**
fr	deb-OPUS-MT	**0.19**	**0.22**	0.02	0.05	**-0.25**	**0.17**	**-0.11**	**0.23**	**0.32**	**0.23**	**0.13**	**0.24**	-0.03	**0.31**	**-0.08**	0.01	**-0.14**	N/A	N/A	**-0.12**
	M2M100	**-0.12**	0.11	**-0.23**	**0.50**	0.09	**0.49**	0.03	-0.03	0.11	**0.21**	0.08	**0.70**	**0.31**	**0.47**	0.05	**-0.08**	**-0.40**	N/A	N/A	0.06
	OPUS-MT	**-0.27**	0.00	**-0.11**	**0.26**	-0.02	**0.19**	**0.13**	-0.11	**0.17**	0.03	**-0.14**	**0.27**	**0.21**	**0.22**	**0.13**	**-0.11**	**-0.23**	N/A	N/A	0.06
uk	deb-OPUS-MT	**-0.43**	-0.06	**-0.12**	**0.41**	0.00	**0.18**	0.07	**-0.24**	0.01	0.01	**-0.17**	**0.23**	**0.17**	**0.16**	**0.10**	**-0.11**	**-0.17**	N/A	N/A	**0.09**
	M2M100	**0.08**	-0.04	**-0.18**	**0.27**	0.02	**0.33**	**0.18**	**0.19**	**0.37**	-0.03	**-0.17**	**0.53**	**0.34**	**0.50**	**0.26**	**-0.30**	**-0.42**	N/A	N/A	**0.11**
	OPUS-MT	0.01	**-0.28**	**-0.41**	0.00	-0.12	0.08	**-0.10**	**-0.19**	**-0.20**	**-0.41**	**-0.39**	**0.07**	-0.03	**0.14**	0.03	**-0.10**	**-0.24**	**-0.40**	0.04	**0.35**
ru	deb-OPUS-MT	**0.23**	**-0.10**	**-0.17**	0.05	0.03	0.05	-0.03	**-0.10**	-0.04	**-0.11**	**-0.14**	**0.09**	**0.04**	**-0.13**	0.03	**-0.08**	**-0.13**	**-0.26**	0.00	**0.13**
	M2M100	**-0.74**	**-0.16**	**-0.21**	0.10	-0.04	**-0.12**	**-0.25**	**0.12**	**0.12**	**-0.20**	**-0.18**	**0.06**	-0.02	**0.08**	0.01	**-0.07**	**-0.11**	**-0.20**	**-0.22**	**0.18**
										SE											
	OPUS-MT	**-1.64**	**0.19**	-0.08	0.13	-0.19	0.09	**-0.29**	-0.01	0.02	0.06	0.07	**0.28**	**-0.12**	**0.17**	**-0.22**	**0.15**	**-0.08**	N/A	N/A	-0.05
es	deb-OPUS-MT	**-0.06**	**0.27**	**0.13**	-0.01	-0.07	0.00	**-0.19**	0.06	**0.10**	**0.16**	**0.12**	-0.01	**-0.18**	-0.01	**-0.19**	**0.20**	**0.08**	N/A	N/A	**-0.20**
	M2M100	**-1.71**	**0.23**	-0.04	0.20	**-0.29**	**0.17**	**-0.16**	0.10	**0.18**	**0.15**	0.08	**0.28**	-0.06	**0.25**	**-0.09**	**0.07**	**-0.16**	N/A	N/A	-0.10
	OPUS-MT	**-1.59**	**0.19**	0.00	0.09	**-0.25**	0.05	**-0.17**	**0.23**	**0.25**	**0.15**	**0.11**	**0.11**	**-0.12**	**0.10**	**-0.15**	**0.14**	-0.04	N/A	N/A	-0.09
fr	deb-OPUS-MT	-0.01	**0.24**	**0.08**	0.11	**-0.17**	0.04	**-0.10**	**0.23**	**0.20**	**0.19**	**0.18**	**-0.06**	**-0.22**	0.01	**-0.15**	**0.18**	0.03	N/A	N/A	**-0.16**
	M2M100	**-0.15**	**0.29**	**0.16**	0.03	-0.05	0.02	**-0.14**	**0.27**	**0.24**	**0.27**	**0.30**	0.01	**-0.14**	0.01	**-0.15**	**0.19**	**0.06**	N/A	N/A	**-0.22**
	OPUS-MT	**-1.54**	0.00	**-0.12**	**0.21**	-0.05	0.10	0.02	-0.09	0.05	0.01	-0.08	**0.13**	0.04	**0.10**	0.02	0.00	**-0.14**	N/A	N/A	0.06
uk	deb-OPUS-MT	**0.09**	**0.12**	**-0.06**	**0.13**	-0.07	**0.08**	0.01	-0.02	**0.08**	**0.15**	**0.09**	**0.06**	**-0.11**	**-0.07**	**-0.10**	**0.13**	**0.06**	N/A	N/A	**0.09**
	M2M100	**-1.72**	-0.04	**-0.14**	0.20	0.03	**0.15**	0.08	0.05	**0.16**	-0.02	**-0.11**	**0.22**	**0.10**	**0.18**	**0.09**	**-0.09**	**-0.19**	N/A	N/A	0.09
	OPUS-MT	**-1.50**	**-0.17**	**-0.25**	**0.21**	0.07	0.00	-0.02	**-0.13**	**-0.18**	**-0.25**	**-0.28**	0.03	-0.03	**0.10**	0.00	**-0.06**	**-0.14**	**-0.32**	-0.02	**0.21**
ru	deb-OPUS-MT	**-0.06**	0.04	0.01	0.11	0.05	0.00	-0.06	**-0.22**	**-0.13**	**0.07**	-0.01	**-0.11**	**-0.15**	**-0.05**	**-0.11**	**0.09**	**0.07**	**-0.24**	**-0.12**	0.03
	M2M100	**-1.57**	-0.09	**-0.14**	**0.11**	0.10	0.04	**-0.18**	-0.01	0.00	**-0.10**	**-0.10**	0.00	-0.06	0.03	-0.01	-0.01	**-0.07**	**-0.12**	**-0.18**	0.11
										GE											
	OPUS-MT	**0.02**	**0.30**	**0.04**	**-0.17**	**0.16**	**-0.10**	**-0.14**	**0.11**	**0.15**	**0.17**	**0.12**	**0.19**	**-0.12**	**0.19**	**-0.20**	**0.15**	**-0.06**	N/A	N/A	**0.17**
es	deb-OPUS-MT	**-0.04**	**0.27**	**0.06**	**0.16**	**-0.10**	**0.15**	**-0.12**	**0.09**	**0.17**	**0.17**	**0.09**	**0.21**	**-0.09**	**0.24**	**-0.16**	**0.09**	**-0.06**	N/A	N/A	**-0.16**
	M2M100	**-0.10**	**0.16**	0.00	**0.08**	**-0.10**	**0.09**	**-0.08**	**0.05**	**0.07**	**0.10**	**0.09**	**0.09**	**-0.08**	**0.08**	**-0.16**	**0.12**	-0.01	N/A	N/A	**-0.08**
	OPUS-MT	0.01	**0.20**	**0.05**	**0.12**	**-0.06**	**0.07**	**-0.11**	**0.09**	**0.13**	**0.12**	**0.08**	**0.09**	**-0.10**	**0.06**	**-0.19**	**0.15**	**0.03**	N/A	N/A	**-0.13**
fr	deb-OPUS-MT	**-0.05**	**0.18**	**0.04**	**-0.11**	**0.08**	**0.08**	**-0.11**	**0.05**	**0.12**	**0.12**	**0.06**	**0.11**	**-0.09**	**0.10**	**-0.17**	**0.11**	0.01	N/A	N/A	**-0.11**
	M2M100	**-0.02**	**0.19**	0.02	**0.12**	**-0.10**	**0.09**	**-0.12**	**0.08**	**0.10**	**0.11**	**0.07**	**0.08**	**-0.11**	**0.04**	**-0.19**	**0.17**	**0.02**	N/A	N/A	**-0.11**
	OPUS-MT	**-0.04**	**0.05**	**-0.04**	**0.07**	-0.03	0.03	**-0.07**	**0.06**	**0.06**	0.02	0.02	0.02	**-0.06**	**-0.11**	**-0.15**	**0.16**	**0.06**	N/A	N/A	0.01
uk	deb-OPUS-MT	**-0.03**	**0.05**	0.02	**0.07**	-0.03	0.00	**-0.06**	**0.06**	**0.07**	**0.03**	**-0.03**	**-0.02**	**-0.05**	**-0.16**	**-0.17**	**0.17**	**0.14**	N/A	N/A	**0.03**
	M2M100	**-0.04**	**0.06**	**-0.04**	**0.09**	-0.05	**0.05**	**-0.06**	**0.07**	**0.08**	0.01	0.01	**0.03**	**-0.07**	**-0.07**	**-0.18**	**0.16**	**0.06**	N/A	N/A	0.01
	OPUS-MT	-0.01	0.00	**-0.03**	-0.02	-0.01	0.00	**-0.04**	0.03	**0.03**	**-0.04**	**-0.06**	-0.01	**-0.04**	**-0.11**	**-0.13**	**0.12**	**0.08**	**-0.04**	0.00	-0.02
ru	deb-OPUS-MT	**-0.02**	-0.01	-0.02	**-0.05**	**-0.05**	-0.01	-0.02	0.01	0.02	**-0.04**	**-0.05**	**-0.03**	**-0.04**	**-0.15**	**-0.15**	**0.13**	**0.12**	**-0.05**	-0.01	-0.02
	M2M100	**0.02**	**0.07**	0.02	-0.01	**-0.06**	-0.01	**-0.07**	**0.09**	**0.07**	0.02	**0.04**	-0.01	**-0.06**	**-0.05**	**-0.11**	**0.10**	**0.06**	**-0.04**	**-0.03**	**0.04**

Table 10: ANOVA results: single effects of bias cues (**F**eminine, **M**asculine, **S**ubject and **O**bject) on norm-$\mathcal{H}$ (S3E), norm-$\mathcal{H}$ (SE) and norm-$\mathcal{H}$ (GE). Values correspond to effect coefficients (deviations from a reference group). **Boldface** indicates statistical significance ($p < 0.05$). The sign of the values indicates whether the presence of the variable increases (positive) or decreases (negative) the mean $\mathcal{H}$ of the group containing the given variable value. Reference group is N for all columns except: 'no name' for Names, 'no default' for Default M, 'unambiguous' for Ambiguity.

Lang.	Model	Names	Recency		Implicit Causality				Stereotype				Subject		Context				Default M		Ambiguity
			F	M	S F	S M	O F	O M	S F	S M	O F	O M	F	M	S F	S M	O F	O M	S	O	
									S3E												
	OPUS-MT	**12.12**	**47.0**	**83.53**	-2.45	**20.67**	-2.44	**27.76**	**38.98**	**49.12**	**44.16**	**37.6**	0.23	**31.36**	**7.95**	**42.89**	**-22.44**	**14.64**	N/A	N/A	**-65.29**
es	deb-OPUS-MT	**-2.35**	**36.52**	**79.64**	**-58.1**	-7.06	**18.24**	**-8.44**	**31.09**	**33.94**	**42.22**	**36.85**	**32.48**	-1.86	**34.19**	**6.06**	**48.72**	**-29.24**	N/A	N/A	**14.06**
	M2M100	**0.19**	**0.54**	0.02	**-0.28**	**0.38**	**-0.27**	**0.37**	**-0.26**	**0.18**	**0.25**	**0.25**	**0.17**	**0.31**	**-0.25**	**0.31**	**-0.29**	**0.31**	N/A	N/A	**-0.17**
	OPUS-MT	**0.12**	**0.41**	**0.06**	**0.32**	**-0.15**	**0.23**	**-0.17**	**0.18**	**0.25**	**0.23**	**0.14**	**0.26**	**-0.16**	**0.32**	**-0.19**	**0.16**	**-0.14**	N/A	N/A	**-0.24**
fr	deb-OPUS-MT	**-0.02**	**0.26**	**0.06**	**-0.16**	**0.23**	**-0.07**	**0.15**	**-0.1**	**0.12**	**0.21**	**0.18**	**0.08**	**0.21**	**-0.08**	**0.25**	**-0.12**	**0.05**	N/A	N/A	**-0.08**
	M2M100	**-0.12**	0.11	**-0.23**	0.06	**0.5**	0.09	**0.49**	0.03	-0.03	0.11	**0.21**	0.08	**0.7**	**0.31**	**0.47**	0.05	**-0.08**	N/A	N/A	**-0.4**
	OPUS-MT	**0.12**	**0.22**	0.03	**0.16**	**-0.07**	**0.09**	**-0.11**	**0.15**	**0.18**	**0.14**	**0.1**	**0.09**	**-0.1**	**0.05**	**-0.1**	**0.16**	**-0.04**	**-0.09**	**-0.05**	**-0.12**
uk	deb-OPUS-MT	0.83	**-10.2**	**24.93**	**-7.38**	**-19.32**	**18.37**	**-14.82**	**27.31**	**5.26**	**9.67**	**8.87**	**5.29**	**-15.13**	**23.85**	**-10.15**	**44.14**	**-26.98**	1.7	-2.21	-0.34
	M2M100	**0.14**	**0.3**	**0.06**	**-0.18**	**0.2**	**-0.16**	**0.12**	**-0.13**	**0.19**	**0.19**	**0.16**	**0.15**	**0.1**	**-0.15**	**0.05**	**-0.17**	**0.22**	**-0.02**	**-0.1**	**-0.04**
	OPUS-MT	**15.38**	**4.07**	**52.95**	**-27.98**	**17.38**	**-24.19**	**20.45**	**11.79**	**12.51**	**12.13**	**13.6**	**-24.94**	**22.11**	**-28.52**	**41.31**	**-18.56**	**23.26**	-0.79	-1.99	**-28.53**
ru	deb-OPUS-MT	**0.23**	**-0.1**	**-0.17**	**0.13**	0.05	0.03	0.05	-0.03	**-0.1**	-0.04	**-0.11**	**-0.14**	**0.09**	**0.04**	**0.13**	0.03	**-0.08**	**-0.13**	**-0.26**	-0.0
	M2M100	**0.08**	**-0.07**	**-0.05**	**0.06**	**-0.02**	-0.01	**-0.02**	-0.0	**-0.06**	**-0.07**	**-0.05**	**-0.03**	**-0.01**	0.0	**-0.02**	0.0	**-0.01**	-0.0	0.01	**0.03**
									SE												
	OPUS-MT	**12.12**	**47.0**	**83.53**	-2.45	**20.67**	-2.44	**27.76**	**38.98**	**49.12**	**44.16**	**37.6**	0.23	**31.36**	**7.95**	**42.89**	**-22.44**	**14.64**	N/A	N/A	**-65.29**
es	deb-OPUS-MT	**-0.08**	**-0.06**	**-0.05**	**0.05**	-0.01	0.0	**-0.01**	-0.0	**-0.04**	**-0.05**	**-0.04**	**-0.03**	**-0.01**	0.0	**-0.01**	**0.01**	**-0.02**	N/A	N/A	**-0.01**
	M2M100	**0.08**	**-0.07**	**-0.06**	**0.07**	-0.01	-0.0	**-0.02**	**-0.01**	**-0.05**	**-0.05**	**-0.04**	**-0.04**	**-0.01**	0.0	**-0.02**	**0.01**	**-0.02**	N/A	N/A	**-0.01**
	OPUS-MT	**16.01**	**47.22**	**80.76**	-0.14	**19.01**	**-8.69**	**26.55**	**31.41**	**35.76**	**33.4**	**31.98**	-1.93	**29.87**	2.67	**35.97**	**-15.81**	**17.83**	N/A	N/A	**-64.01**
fr	deb-OPUS-MT	**-0.07**	**-0.05**	**-0.04**	**0.04**	-0.0	**0.02**	-0.01	0.0	**-0.03**	**-0.04**	**-0.03**	**-0.03**	**-0.01**	**0.01**	**-0.01**	**0.01**	**-0.02**	N/A	N/A	**-0.01**
	M2M100	**0.14**	**0.38**	**0.04**	**-0.21**	**0.35**	**-0.15**	**0.23**	**-0.17**	**0.15**	**0.2**	**0.19**	**0.13**	**0.23**	**-0.16**	**0.24**	**-0.19**	**0.19**	N/A	N/A	**-0.12**
	OPUS-MT	**-0.01**	**-0.01**	**-0.01**	**-0.03**	-0.02	0.0	0.0	**-0.03**	**-0.03**	**-0.01**	-0.01	-0.0	0.0	**-0.01**	0.01	-0.0	-0.0	**0.02**	**0.03**	**0.01**
uk	deb-OPUS-MT	0.0	**0.09**	0.02	**-0.05**	0.01	-0.06	**0.06**	-0.04	**0.12**	**0.16**	**0.08**	**0.04**	**0.04**	**-0.02**	**0.08**	0.02	0.01	**-0.06**	**-0.1**	**-0.03**
	M2M100	**0.14**	**0.3**	**0.06**	**-0.18**	**0.2**	**-0.16**	**0.12**	**-0.13**	**0.19**	**0.19**	**0.16**	**0.15**	**0.1**	**-0.15**	**0.05**	**-0.17**	**0.22**	**-0.02**	**-0.1**	**-0.04**
	OPUS-MT	**15.38**	**4.07**	**52.95**	**-27.98**	**17.38**	**-24.19**	**20.45**	**11.79**	**12.51**	**12.13**	**13.6**	**-24.94**	**22.11**	**-28.52**	**41.31**	**-18.56**	**23.26**	-0.79	-1.99	**-28.53**
ru	deb-OPUS-MT	1.34	-3.62	**46.99**	**-21.71**	**-33.31**	**16.73**	**-27.33**	**19.72**	**9.4**	**7.79**	**6.6**	**9.66**	**-28.51**	**22.12**	**-24.71**	**50.74**	**-26.37**	**15.86**	-0.28	-0.68
	M2M100	**20.01**	**18.76**	**50.38**	**-34.59**	**-11.85**	5.91	**-16.63**	**13.63**	**14.94**	**15.4**	**14.21**	**15.2**	**-17.46**	**14.68**	**-15.4**	**30.42**	**-10.35**	**16.48**	-0.74	-3.23
									GE												
	OPUS-MT	**-0.0**	**-0.02**	**-0.01**	-0.01	0.01	**-0.01**	0.0	**-0.02**	**-0.02**	**-0.01**	**-0.01**	**-0.01**	**0.01**	-0.0	**0.01**	**-0.01**	-0.0	N/A	N/A	**0.02**
es	deb-OPUS-MT	**-2.35**	**36.52**	**79.64**	**-58.1**	-7.06	**18.24**	**-8.44**	**31.09**	**33.94**	**42.22**	**36.85**	**32.48**	-1.86	**34.19**	**6.06**	**48.72**	**-29.24**	N/A	N/A	**14.06**
	M2M100	**0.19**	**0.54**	0.02	**-0.28**	**0.38**	**-0.27**	**0.37**	**-0.26**	**0.18**	**0.25**	**0.25**	**0.17**	**0.31**	**-0.25**	**0.31**	**-0.29**	**0.31**	N/A	N/A	**-0.17**
	OPUS-MT	**0.06**	**-0.05**	**-0.04**	0.01	**0.02**	**-0.01**	0.01	**-0.04**	**-0.04**	**-0.03**	**-0.02**	**-0.0**	**0.01**	**-0.01**	**0.01**	**-0.02**	**-0.01**	N/A	N/A	**0.04**
fr	deb-OPUS-MT	**-0.05**	**0.18**	**0.04**	**-0.11**	**0.08**	**-0.11**	**0.08**	**-0.11**	**0.05**	**0.12**	**0.12**	**0.06**	**0.11**	**-0.09**	**0.1**	**-0.17**	**0.11**	N/A	N/A	0.01
	M2M100	**0.14**	**0.38**	**0.04**	**-0.21**	**0.35**	**-0.15**	**0.23**	**-0.17**	**0.15**	**0.2**	**0.19**	**0.13**	**0.23**	**-0.16**	**0.24**	**-0.19**	**0.19**	N/A	N/A	**-0.12**
	OPUS-MT	**0.12**	**0.22**	0.03	**0.16**	**-0.07**	**0.09**	**-0.11**	**0.15**	**0.18**	**0.14**	**0.1**	**0.09**	**-0.1**	**0.05**	**-0.1**	**0.16**	**-0.04**	**-0.09**	**-0.05**	**-0.12**
uk	deb-OPUS-MT	0.0	**0.09**	0.02	**-0.05**	0.01	-0.06	**0.06**	-0.04	**0.12**	**0.16**	**0.08**	**0.04**	**0.04**	**-0.02**	**0.08**	0.02	0.01	**-0.06**	**-0.1**	**-0.03**
	M2M100	**0.14**	**0.3**	**0.06**	**-0.18**	**0.2**	**-0.16**	**0.12**	**-0.13**	**0.19**	**0.19**	**0.16**	**0.15**	**0.1**	**-0.15**	**0.05**	**-0.17**	**0.22**	**-0.02**	**-0.1**	**-0.04**
	OPUS-MT	**0.09**	**0.17**	**0.05**	0.01	**-0.11**	**0.05**	**-0.08**	**0.08**	**0.13**	**0.11**	**0.06**	**0.04**	**-0.08**	-0.02	**-0.06**	**0.14**	-0.01	**-0.08**	**-0.03**	**-0.11**
ru	deb-OPUS-MT	**-0.02**	-0.01	-0.02	0.02	**-0.05**	**-0.05**	-0.01	-0.02	0.01	0.02	**-0.04**	**-0.05**	**-0.03**	**-0.04**	**-0.15**	**-0.15**	**0.13**	**0.12**	**-0.05**	-0.01
	M2M100	**0.15**	**0.29**	**0.08**	**-0.18**	**0.13**	**-0.1**	**0.09**	**-0.12**	**0.16**	**0.21**	**0.15**	**0.11**	**0.07**	**-0.13**	**0.08**	**-0.05**	**0.15**	**-0.08**	**-0.1**	-0.02

Table 11: ANOVA results (no normalisation): single effects of bias cues (**F**eminine, **M**asculine, **S**ubject and **O**bject) on $\mathcal{H}$ (S3E), $\mathcal{H}$ (SE) and $\mathcal{H}$ (GE). Values correspond to effect coefficients (deviations from a reference group). **Boldface** indicates statistical significance ($p < 0.05$). The sign of the values indicates whether the presence of the variable increases (positive) or decreases (negative) the mean $\mathcal{H}$ of the group containing the given variable value. Reference group is N for all columns except: 'no name' for Names, 'no default' for Default M, 'unambiguous' for Ambiguity.

H Gender Accuracy

Table 12 presents more fine-grained results than Table 3 with regard to gender accuracy, namely splitting the results by subset of the dataset. The results in the Ambiguous column are not meaningfully interpretable, as a single ground truth label of gender cannot capture the true desired behavior of the model, especially when the gold label for ambiguous cases is mostly 'neutral', and neutral is not commonly used as grammatical gender for animate objects in the languages used in this study. The case of Russian, where the performance increases on the Ambiguous subset actually reflects the model choosing the masculine forms, which are tagged as 'neutral' by the morphological parser due to the masculine form often being the default choice for both genders, as discussed in Section 4.

Lang.	Model	All	Pro	Anti	Unamb.	Amb.
es	OPUS-MT	55.20	67.95	52.10	67.95	33.96
	deb-OPUS-MT	55.69	68.13	52.95	68.13	34.39
	M2M100	55.68	70.77	51.17	70.77	32.40
	TOWER+9B	79.44	86.91	88.08	86.91	44.02
fr	OPUS-MT	52.05	64.27	46.55	64.27	37.25
	deb-OPUS-MT	52.98	64.79	48.10	64.79	37.75
	M2M100	50.95	61.66	47.57	61.66	34.84
	TOWER+9B	69.96	75.77	75.13	75.77	45.78
uk	OPUS-MT	38.65	45.34	34.20	45.34	33.75
	deb-OPUS-MT	38.95	46.12	34.15	46.12	33.74
	M2M100	40.97	47.76	36.81	47.76	35.20
	TOWER+9B	55.05	57.58	58.97	57.58	40.88
ru	OPUS-MT	39.50	48.57	33.27	48.57	33.24
	deb-OPUS-MT	39.50	48.42	33.38	48.42	33.33
	M2M100	41.01	48.49	36.81	48.49	33.81
	TOWER+9B	49.50	51.44	52.68	51.44	48.25

Table 12: Comparison of Gender Accuracy Overall, in Pro-/Anti-Stereotypical and Ambiguous Cases Across Models, on WINOMT.

Lang.	Model	All	Pro	Anti	Unamb.	Amb.
es	OPUS-MT	81.35	85.80	83.37	84.55	75.14
	deb-OPUS-MT	81.31	85.62	83.43	84.49	75.14
	M2M100	79.56	84.44	81.29	82.82	73.25
	TOWER+9B	81.95	85.79	84.21	85.00	76.08
fr	OPUS-MT	77.63	82.23	81.16	81.66	70.47
	deb-OPUS-MT	77.69	82.41	80.88	81.60	70.73
	M2M100	76.24	80.69	78.65	79.61	70.25
	TOWER+9B	81.51	85.63	83.76	84.70	75.95
uk	OPUS-MT	80.56	85.57	82.73	84.10	73.69
	deb-OPUS-MT	80.19	84.85	82.13	83.45	73.86
	M2M100	81.27	85.89	84.09	84.96	74.10
	TOWER+9B	85.51	90.18	88.05	89.12	78.57
ru	OPUS-MT	82.53	86.20	84.99	85.58	76.86
	deb-OPUS-MT	82.76	86.46	85.27	85.85	77.02
	M2M100	81.71	86.39	84.06	85.21	75.21
	TOWER+9B	85.06	89.46	87.27	88.37	78.96

Table 13: Comparison of COMET Scores Overall, in Pro-/Anti-Stereotypical and Ambiguous Cases Across Models, on the 100 manually translated sentences.

I Quality Estimation with Human Translations

Table 13 presents the results of the models used in this study on the 100 human-annotated instances. For unambiguous cases we use a single reference, whereas for ambiguous cases, we calculate performance by taking the maximum COMET score of both acceptable translations.

Language	Model	LogProb (Correct)	LogProb (Incorrect)	S3E I (Correct)	S3E I (Incorrect)	SE I (Correct)	SE I (Incorrect)	GE I (Correct)	GE I (Incorrect)
ES	OPUS-MT	-149.7	-149.78	7.83	8.88	0.3	0.35	0.33	0.3
	deb-OPUS-MT	-149.19	-149.01	8.08	9.16	0.29	0.31	0.35	0.33
	M2M100	-226.61	-227.29	23.61	26.03	0.41	0.4	0.42	0.43
	TOWER+9B	-586.29	-586.72	15.25	19.80	0.97	1.19	0.51	0.51
FR	OPUS-MT	-197.1	-195.11	9.18	9.89	0.73	0.72	0.24	0.29
	deb-OPUS-MT	-196.98	-195.09	9.18	9.85	0.48	0.52	0.26	0.33
	M2M100	-283.91	-281.71	186.42	194.52	0.49	0.4	0.43	0.48
	TOWER+9B	-661.37	-677.52	15.62	19.79	1.24	1.33	0.78	0.78
UK	OPUS-MT	-161.98	-161.14	147.6	152.15	0.6	0.54	0.22	0.22
	OPUS-MT-debiased	-161.46	-160.68	150.52	153.76	0.49	0.47	0.23	0.23
	M2M100	-241.0	-241.49	204.72	211.9	0.28	0.24	0.23	0.25
	TOWER+9B	-455.91	-462.70	16.84	18.81	1.5	1.59	1.62	1.47
RU	OPUS-MT	-170.72	-170.9	32.14	33.15	0.38	0.43	0.08	0.19
	deb-OPUS-MT	-170.58	-170.75	32.31	33.27	0.32	0.37	0.06	0.17
	M2M100	-220.25	-220.78	218.11	219.06	0.45	0.4	0.16	0.3
	TOWER+9B	-499.57	-508.94	27.34	29.75	1.95	2.11	1.0	0.78

Table 14: Log Probability and Surprisal Measures across Models and Languages

J Log Probability and Surprisal Scores

Table 14 presents the Log probability and surprisal scores, as well as their relative differences between the Correct and Incorrect translations of the Unambiguous instances in WINOMT.

K Results on the MGENTE dataset

Table 15 presents the results of the experiments on the MGENTE dataset.

Lang.	Model	Gender Acc	ΔLog prob	ΔI (S3E)	COMET
ES	OPUS-MT	60.40	-0.06	-1.92	84.90
	deb-OPUS-MT	60.67	-0.06	-1.91	84.86
	M2M100	60.53	-0.05	1.73	84.53
	TOWER+9B	67.47	-0.04	4.53	86.88
DE	OPUS-MT	69.20	-0.06	-1.45	81.30
	deb-OPUS-MT	68.93	-0.06	-1.43	81.24
	M2M100	68.53	-0.06	1.96	82.29
	TOWER+9B	79.87	-0.04	2.21	85.33
IT	OPUS-MT	64.40	-0.04	-1.45	84.65
	deb-OPUS-MT	64.40	-0.04	-1.46	84.58
	M2M100	66.00	-0.04	1.96	85.04
	TOWER+9B	74.00	-0.03	2.21	87.78
EL	OPUS-MT	65.73	-0.05	-1.43	78.41
	deb-OPUS-MT	65.87	-0.05	-1.41	77.90
	M2M100	64.27	-0.08	1.96	78.58
	TOWER+9B	64.11	-0.02	2.18	66.23

Table 15: Gender Accuracy, Δ Log Probability, and ΔI (S3E) on Unambiguous instances of MGENTE, COMET scores on WMT test sets.

There is no correlation between ΔI scores and gender accuracy on the MGENTE dataset (see Table 16).

Table 17 presents the entropy (S3E) scores on the MGENTE dataset. The best performing model (TOWER+9B) shows the lowest $\Delta\mathcal{H}$ across languages, indicating the least bias.

L Rankings by Different Metrics

Table 18 presents the rankings of models according to the metrics employed in this study.

Correlation	Metric	Statistic	p-value
Spearman	ΔI (S3E)	-0.40	0.12
	ΔI (SE)	-0.35	0.19
	ΔI (GE)	0.44	0.09
	ΔLog prob	-0.11	0.70
Kendall	ΔI (S3E)	-0.28	0.14
	ΔI (SE)	-0.18	0.34
	ΔI (GE)	0.30	0.11
	ΔLog prob	-0.10	0.59

Table 16: Spearman and Kendall correlations between ΔI under different uncertainty metrics and Log Probabilities on the one hand, and gender accuracy on the other on the MGENTE dataset. No statistically significant correlations present.

M Rank Correlation

Table 19 presents the correlation scores between ΔI and Δ Log Probabilies on the one hand, and gender accuracy scores on the other.

Lang.	Model	Unamb	Amb	$\Delta\mathcal{H}$
ES	OPUS-MT	2.53	2.57	**-0.02**
	deb-OPUS-MT	2.45	2.47	**-0.01**
	M2M100	2.41	2.28	0.06
	TOWER+9B	3.77	4.08	**-0.08**
DE	OPUS-MT	2.90	2.87	0.01
	deb-OPUS-MT	2.92	2.90	0.01
	M2M100	2.72	2.90	-0.06
	TOWER+9B	3.36	4.35	**-0.25**
IT	OPUS-MT	2.89	2.78	0.04
	deb-OPUS-MT	2.17	1.79	0.19
	M2M100	2.91	2.63	0.10
	TOWER+9B	3.77	4.08	**-0.08**
EL	OPUS-MT	2.80	2.67	0.05
	deb-OPUS-MT	2.75	2.60	0.05
	M2M100	3.19	2.88	0.10
	TOWER+9B	6.74	6.99	-0.04

Table 17: Unambiguous and Ambiguous $\mathcal{H}$ (S3E) on MGENTE. Statistically significant correlations between the $\Delta\mathcal{H}$ and COMET scores at the instance level are presented in **bold**.

N Entropy Scores

Table 20 presents the $\mathcal{H}$ scores and their relative differences between the unambiguous and ambiguous settings for different UQ metrics used in this study.

Gender Acc	**Delta S**
TOWER-ES	TOWER-ES
M2M100-ES	TOWER-FR
TOWER-FR	deb-OPUS-MT-ES
deb-OPUS-MT-ES	M2M100-ES
OPUS-MT-ES	TOWER-UK
deb-OPUS-MT-FR	OPUS-MT-ES
OPUS-MT-FR	TOWER-RU
M2M100-FR	deb-OPUS-MT-FR
TOWER-UK	M2M100-FR
TOWER-RU	OPUS-MT-FR
OPUS-MT-UK	OPUS-MT-RU
M2M100-UK	deb-OPUS-MT-UK
deb-OPUS-MT-UK	deb-OPUS-MT-RU
M2M100-RU	M2M100-UK
deb-OPUS-MT-RU	M2M100-RU
OPUS-MT-RU	OPUS-MT-UK

Table 18: Model rankings for gender accuracy (higher is better) and S3E delta (lower is better) on the WINOMT dataset.

Correlation	**Metric**	**Statistic**	**p-value**
Spearman	ΔI (S3E)	**-0.84**	**0.00**
	ΔI (SE)	-0.42	0.10
	ΔI (GE)	0.29	0.27
	ΔLog prob	0.00	1.00
Kendall	ΔI (S3E)	**-0.63**	**0.00**
	ΔI (SE)	-0.30	0.12
	ΔI (GE)	0.21	0.27
	ΔLog prob	-0.03	0.89

Table 19: Spearman and Kendall correlations between ΔI under different uncertainty metrics and Log Probabilities on the one hand, and gender accuracy on the other, for the WINOMT dataset. Statistically significant correlations ($p < 0.05$) are in **bold**.

Language	Model	S3E			SE			GE		
		Unamb.	Amb.	$\Delta\mathcal{H}$	Unamb.	Amb.	$\Delta\mathcal{H}$	Unamb.	Amb.	$\Delta\mathcal{H}$
ES	OPUS-MT	1.23	1.12	0.09	0.33	0.22	0.33	0.21	0.13	0.38
	deb-OPUS-MT	0.97	0.89	0.08	0.41	0.25	0.39	0.24	0.16	0.33
	M2M100	1.79	1.45	0.19	0.46	0.17	0.63	0.25	0.09	0.64
FR	OPUS-MT	1.79	1.43	0.20	0.57	0.40	0.30	0.23	0.08	0.65
	deb-OPUS-MT	1.21	1.08	0.11	0.64	0.50	0.22	0.23	0.09	0.61
	M2M100	3.22	2.78	0.14	0.56	0.29	0.48	0.25	0.17	0.32
UK	OPUS-MT	1.96	2.16	-0.10	0.40	0.39	0.03	0.20	0.14	0.30
	deb-OPUS-MT	1.98	2.15	-0.09	0.44	0.41	0.07	0.22	0.15	0.32
	M2M100	2.05	2.28	-0.11	0.37	0.41	-0.11	0.18	0.16	0.11
RU	OPUS-MT	1.56	1.68	-0.08	0.38	0.32	0.16	0.12	0.03	0.75
	deb-OPUS-MT	1.05	0.97	0.08	0.43	0.42	0.02	0.12	0.06	0.50
	M2M100	1.83	2.29	-0.25	0.50	0.29	0.42	0.15	0.10	0.33

Table 20: $\mathcal{H}$ scores across models and languages, with relative differences ($\Delta\mathcal{H}$) between unambiguous and ambiguous conditions on the WINOMT dataset.