Watching the Watchers: Exposing Gender Disparities in Machine Translation Quality Estimation

Emmanouil Zaranis[*♣♠], **Giuseppe Attanasio**[*♣], **Sweta Agrawal**[♣], **André F.T. Martins**[♣♠◇]
[♣]Instituto de Telecomunicações, Lisbon, Portugal
[♠]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[◇]Unbabel, Lisbon, Portugal
emmanouil.zaranis@tecnico.ulisboa.pt

Abstract

Quality estimation (QE)—the automatic assessment of translation quality—has recently become crucial across several stages of the translation pipeline, from data curation to training and decoding. While QE metrics have been optimized to align with human judgments, whether they encode social biases has been largely overlooked. Biased QE risks favoring certain demographic groups over others, e.g., by exacerbating gaps in visibility and usability. This paper defines and investigates gender bias of QE metrics and discusses its downstream implications for machine translation (MT). Experiments with state-of-the-art QE metrics across multiple domains, datasets, and languages reveal significant bias. When a human entity's gender in the source is undisclosed, masculine-inflected translations score higher than feminine-inflected ones, and gender-neutral translations are penalized. Even when contextual cues disambiguate gender, using context-aware QE metrics leads to more errors in selecting the correct translation inflection for feminine referents than for masculine ones. Moreover, a biased QE metric affects data filtering and quality-aware decoding. Our findings underscore the need for a renewed focus on developing and evaluating QE metrics centered on gender.[1]

1 Introduction

Quality estimation—the automatic evaluation of machine-translated content without human-written references (Callison-Burch et al., 2012)—has gained increasing interest in the natural language processing (NLP) and machine translation communities. Recent work has focused on building QE metrics aligned with human quality judgments using neural encoders (Rei et al., 2020; Perrella et al., 2022; Rei et al., 2023; Juraska et al., 2023; Guerreiro et al., 2024) or autoregressive language models (Kocmi and Federmann, 2023b; Fernandes et al., 2023; Kocmi and Federmann, 2023a; Xu et al., 2023), and has explored their use throughout the translation pipeline, e.g., for data filtering (Peter et al., 2023; Alves et al., 2024), training (Ramos et al., 2024a; Yan et al., 2023; He et al., 2024), and improving output quality (Freitag et al., 2022; Fernandes et al., 2022; Farinhas et al., 2023; Ramos et al., 2024b; Vernikos and Popescu-Belis, 2024).

Evidence has shown QE scores effectively measure translation quality aspects like adequacy and fluency (Guerreiro et al., 2024). However, little is known about whether exogenous factors, e.g., conformity to social norms, play a role. If QE metrics favor normative language, they may penalize non-conforming expressions. Gender norms exemplify this issue—gender-based discrimination can amplify certain groups' visibility in favor of others, leading to representational and allocative harms (Crawford, 2017). For instance, biased metrics may systematically retain masculine-inflected data in parallel corpus filtering while discarding equivalent-quality feminine translations. Similarly, when using QE to improve MT systems (Fernandes et al., 2022), such metrics could lead to systems that predominantly generate masculine translations, forcing women users to spend more time and effort correcting gender errors (Savoldi et al., 2024).

In this paper, we conceptualize and measure gender bias in QE metrics and its downstream effects. We draw two conditions to identify a biased metric. A QE metric exhibits gender bias if, when translating human entities, it systematically (*i*) assigns higher scores to one gender form over others when no explicit lexical cues disclose gender, or (*ii*) fails to detect the correct gender form even when lexical cues are provided, disproportionately across genders. We investigate bias across commonly used QE metrics through an extensive empir-

[1]We release code and artifacts at: `https://github.com/deep-spin/gender-bias-qe-metrics`.
[*]Equal contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25261–25284
July 27 - August 1, 2025

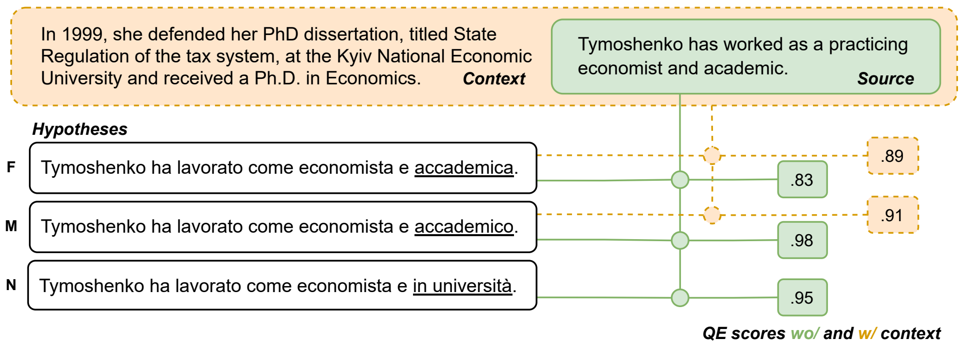

Figure 1: **Our contrastive setup.** MT-GenEval (Currey et al., 2022) sample of an English source sentence (solid green box), its preceding context (dashed orange box), and three minimal-edit Italian translations (bottom left) evaluated by xCOMET XL (Guerreiro et al., 2024). Based on the entity's (Tymoshenko) gender, the noun "academic" can inflect into a feminine (F) or masculine (M) form, or be rephrased with gender-neutral (N) language. QE scores without context (solid) for all three forms and for the F/M (dashed) when context disambiguating gender is considered. The masculine translation scores higher, regardless of whether the context is provided or not.

ical study. We focus on out-of-English translation where the target language uses grammatical gender,[2] a scenario frequently used in prior research on gender bias in MT (e.g., Vanmassenhove et al., 2018; Stanovsky et al., 2019; Saunders et al., 2020; Attanasio et al., 2023; Sant et al., 2024). We address both binary (i.e., contrasting masculine and feminine translations) and gender-neutral (i.e., contrasting gendered and gender-neutral forms) translation scenarios, and experiment with 11 state-of-the-art metrics (supervised and prompted) across two domains, three datasets, and eight target languages.

Our analysis reveals evidence of gender bias across metrics, languages, and experimental conditions. Masculine forms receive higher scores when the source gender is ambiguous. Moreover, metrics favor gendered over neutral expressions. Even with disambiguating context, metrics show higher error rates in identifying feminine gender forms. These biases significantly impact MT pipelines: feminine-inflected translations fail quality thresholds more frequently than masculine equivalents despite equal validity. During inference, QE-based hypothesis reranking amplifies gender bias in MT while improving translation quality. These findings call for extending QE metrics evaluation beyond human judgment alignment to consider gender equity.

2 Related Work

2.1 Evaluating MT Metrics

Significant efforts have been made to evaluate automatic metrics for MT. Most have focused on comparing sentence- or system-level correlations between a metric's score and human judgements (Macháček and Bojar, 2014; Stanojević et al., 2015; Kocmi et al., 2021; Freitag et al., 2023; Deutsch et al., 2023; Thompson et al., 2024). Finer-grained analyses have discovered poor robustness to low-quality translations (Fomicheva and Specia, 2019), domain (Zouhar et al., 2024), or passive voice (Avramidis et al., 2023). Metrics have also been shown to produce imbalanced score distributions (Sun et al., 2020), and fail to distinguish discrepancies in numbers and named entities (Amrhein and Sennrich, 2022) or high-quality translations (Agrawal et al., 2024b). However, none of the studies have thus far evaluated gender bias exhibited by QE metrics.

2.2 Gender Bias in MT

Gender bias in MT systems—stemming from societal norms, model design, and deployment decisions—has been extensively documented (Savoldi et al., 2021). Prior work in NLP has primarily focused on measuring the extent of such bias in system outputs (e.g., Stanovsky et al., 2019; Cho et al., 2019; Hovy et al., 2020; Vanmassenhove and Monti, 2021; Levy et al., 2021; Lucy and Bamman, 2021; Sant et al., 2024; Piergentili et al., 2023; Lauscher et al., 2023; Lardelli et al., 2024; Stewart and Mihalcea, 2024), as well as on proposing mitigation strategies (e.g., Escudé Font and Costa-jussà, 2019; Costa-jussà and de Jorge, 2020; Saunders et al., 2020; Saunders and Byrne, 2020; Escolano et al., 2021; Lee et al., 2023; Garg et al., 2024). In contrast to this line of work, we shift the focus from the translation systems them-

[2]I.e., where gender for human referents is assigned on a semantic basis and reflected with morphological marks.

selves to the task of quality estimation, a step that, typically, follows the MT. Further, we investigate whether bias in QE can propagate to or reinforce bias in MT outputs.

2.3 Social Biases in NLG Metrics

Several studies have investigated whether metrics used to evaluate natural language generation (NLG) exhibit social biases. Qiu et al. (2023) test whether n-gram- and model-based metrics, *e.g.* CLIPScore (Hessel et al., 2021) show gender bias in image captioning. Sun et al. (2022) quantify several social biases. They use English coreference resolution examples to measure the gender bias of several n-grams and model-based NLG metrics, including BERTScore (Zhang* et al., 2020), MoverScore (Zhao et al., 2019), and BLEURT (Sellam et al., 2020). Gao and Wan (2022) measure race and gender stereotypes using WEAT (Caliskan et al., 2017) and SEAT (May et al., 2019). However, no study targets QE metrics for MT or reference-free NLG metrics.

3 Measuring Gender Bias in QE

Several design choices arise when testing how neural QE metrics handle gender-related translations, including (*i*) source and target languages, (*ii*) domain, (*iii*) granularity (document, passage, or sentence-level MT), (*iv*) gender ambiguity in the source, and (*v*) gender-inflected words in the target. In this work, we experiment with **out-of-English** (En→*) MT setups where the target language (*) is characterized by grammatical gender. In these languages, masculine and feminine genders are represented through word choices and inflection (e.g., different suffixes, as in Figure 1), while neopronouns, neo-morphemes, or rephrasing permit beyond-the-binary or neutral translation (Lauscher et al., 2023; Piergentili et al., 2023; Lardelli et al., 2024; Piergentili et al., 2024). Following prominent work on gender bias in MT (Currey et al., 2022; Rarrick et al., 2023), we focus on **sentence-level** translations from **multiple domains** and observe gender-inflected words referring to **human entities**. We challenge QE metrics with translation pairs where the entity's gender in the source is ambiguous or contextual cues disambiguate it.

3.1 Formalizing Quality Estimation

For the remainder of the paper, we will discuss QE metrics that support quality assessment of translations (Graham et al., 2016; Burchardt, 2013). A QE metric is thus a function f such that:

$$\mathrm{QE}(s, h) := f_\theta(s, h) = a, \tag{1}$$

where s is the English source, h the translation to evaluate, and $a \in \mathbb{R}$ is the assessment score. Typically, f is a neural function parameterized by θ, and a is bounded in a range with the edges indicating the poorest and highest quality.

QE Metrics We test 11 QE metrics, varying model sizes and architectures. We include seven neural metrics that ranked highest in recent editions of the WMT QE Shared Task (Blain et al., 2023): CometKiwi 22 (Rei et al., 2022), CometKiwi 23 XL/XXL (Rei et al., 2023), xCOMET XL/XXL (Guerreiro et al., 2024), and MetricX 23 L/XL (Juraska et al., 2023). Following Kocmi and Federmann (2023b), we test the GEMBA-DA prompt with three decoder-only, instruction-tuned, open-weight language models—Mistral 7B (Jiang et al., 2023), Gemma 2 9B (Team et al., 2024), Llama 3.1 70B (Dubey et al., 2024)—and a commercial API model, GPT 4o (Achiam et al., 2023).

Our analysis will investigate whether metrics can use extra-sentential contextual cues (§3.2.3). However, contemporary QE metrics are not trained on nor support context beyond the sentence level off-the-shelf. Thus, we rely on two inference-time strategies to incorporate extra-sentential context (Vernikos et al., 2022): 1) We concatenate the context string c to s and h, i.e., $\mathrm{QE}(c; s, c; h)$, and 2) we apply the same strategy as (1) but automatically translate the context c for the hypothesis, i.e., $\mathrm{QE}(c; s, \mathrm{translate}(c); h)$.[3] The latter variant prevents the creation of a cross-lingual hypothesis, out-of-distribution for the QE metric. We modify the prompt for GEMBA metrics to account for context (full details in Appendix A.1).

3.2 Experimental Conditions

In §1, we defined a biased QE metric as one that exhibits systematic undue preferences for gender forms in the target or disproportionate error rate across gender groups. Here, we describe the datasets and measurement methods we use to assess the presence and extent of gender bias.

3.2.1 Datasets with Minimal Edits

We study datasets organized in **minimal edit** contrastive pairs, where sources and **reference translations** differ only in gender-related words. This

[3] We use NLLB-200 3.3B (NLLB Team et al., 2022) for translating c as it supports all studied languages.

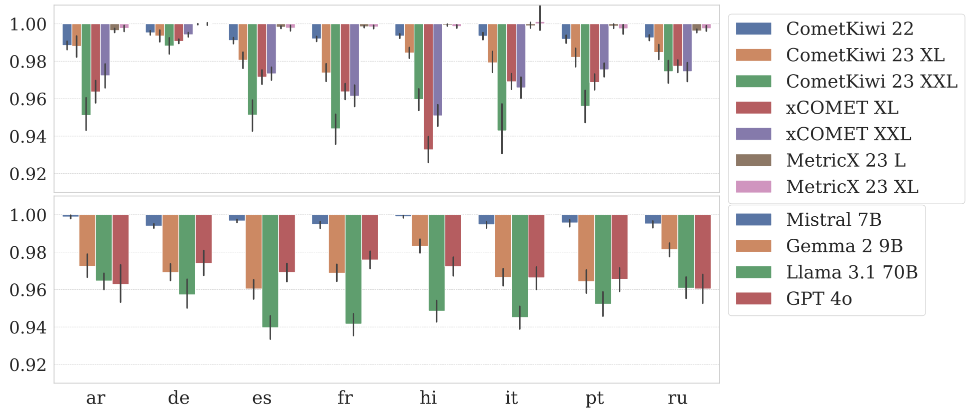

Figure 2: **Ratio** $\mathrm{QE}(s, h_F)/\ \mathrm{QE}(s, h_M)$ **on MT-GenEval, ambiguous instances**. Average and 95% confidence interval on the test set by language. Neural (top) and GEMBA (bottom) QE metrics.

approach has two main advantages. First, it enables precise isolation of effects related to gender-inflected elements (see, for example, translations in Figure 1). Therefore, any change in the QE score can only be attributed to the metric's sensitivity to those words. Second, using human-written references neutralizes potential confounding quality-related effects.[4]

3.2.2 Gender-Ambiguous Entities

With gender ambiguity in the source, we set out to study two contrastive setups: masculine (m) vs. feminine (f) and gendered vs. gender-neutral translations. For the former (m/f) experiments, we use MT-GenEval (Currey et al., 2022) and GATE (Rarrick et al., 2023), two established sentence-level corpora for gender bias evaluation in MT with translations provided by professional translators. From MT-GenEval, we retrieve all test examples from the *contextual* subset, discarding the context that would disambiguate the gender. From GATE, we collect all instances with one gender-marked entity. The two corpora provide linguistic diversity: MT-GenEval collects sentences from Wikipedia, while sources in GATE are written by linguists and cover diverse linguistic phenomena. MT-GenEval includes eight target languages (ar, de, es, fr, hi, it, pt, ru) and GATE three (es, fr, it). Following prior work (Beutel et al., 2019; Gaut et al., 2020; Attanasio et al., 2024), we observe gaps in relativistic terms. We measure the ratio $\mathrm{QE}(s, h_F)/\ \mathrm{QE}(s, h_M)$ where h_F and h_M are the feminine- and masculine-inflected translations, respectively. Echoing the desiderata of counterfactual fairness (Czarnowska et al., 2021), the two scores should not differ, i.e., the ideal ratio is 1.

We use mGeNTE (Savoldi et al., 2025) for gender-neutral experiments. The benchmark targets gender-neutral translations of English sources written by professional translators for it, es, de.[5] We focus on ambiguous sources (Set-N), and measure $\mathrm{QE}(s, h_N)/\ \mathrm{QE}(s, h_G)$, the ratio between the gender-neutral and gendered score. Since gender-neutral translations provide inclusive alternatives that maintain semantic accuracy, QE metrics should assign them higher scores than their gendered counterparts, i.e., the ratio should ideally exceed 1. However, equality between scores represents a minimum acceptable threshold.

3.2.3 Gender-Unambiguous Entities

We study cases where intra- or extra-sentential lexical cues disambiguate gender. For the intra-sentential case, we use the MT-GenEval's *counterfactual* subset. It consists of paired instances with masculine- and feminine-inflected sources, each with its corresponding reference. For the extra-sentential case, we use the MT-GenEval's *contextual* subset, where each test instance is paired with a preceding sentence disambiguating the gender. Figure 1 reports an example of the extra-sentential case. An instance with intra-sentential cues would be: "In her thirties, Tymoshenko has worked as a practicing economist and academic."

Since a single correct gender form exists, given a set of instances S, we compute the error rate (ER)—the fraction of instances where the QE metric as-

[4]Sentence-level outputs are converging toward human-level translation quality in MT research; thus, using references is a valid choice. However, we relax this constraint in §5.2.

[5]Neutralization strategies might extend beyond simple edits. While breaking the minimal-edits requirement, these instances highlight meaningful contrasts between common gendered translations and less common neutral ones.

Metric	ER	Φ
CometKiwi 22	0.11	1.70
CometKiwi 23 XL	0.09	1.18
CometKiwi 23 XXL	**0.07**	**0.87**
xCOMET XL	0.10	1.81
xCOMET XXL	0.08	1.32
MetricX 23 L	0.31	1.25
MetricX 23 XL	0.12	1.19
Llama 3.1 70B	0.31	1.16
Gemma 2 9B	0.28	1.36
Mistral 7B	0.74	**1.13**
GPT 4o	**0.16**	1.15

Table 1: **Total error rate** (ER $\downarrow$) **and error rate ratio between gender groups** ($\Phi(S^F, S^M) \rightarrow 1$). MT-GenEval *counterfactual* set, unambiguous (intra-sentential). Best metrics per type in bold, best overall, underlined. Mean results across eight languages.

signs a higher score to the incorrect gender form. We compute it for all instances and separately for feminine and masculine sources, i.e., $\mathrm{ER}(S^F)$ and $\mathrm{ER}(S^M)$, respectively.[6] To measure gender disparities, we adopt a Multi-group Comparison Metric (Czarnowska et al., 2021) defined as:

$$\Phi(S^F, S^M) = \mathrm{ER}(S^F) / \mathrm{ER}(S^M) \qquad (2)$$

Loosely inspired by False Positive Equality Difference (Dixon et al., 2018), the metric operationalizes the statement: "Even when lexical cues leave no doubts on gender identity, sources with feminine referents exhibit Φ times the number of errors in gender inflection as sources with masculine referents." The ideal scenario is equal errors, i.e., a ratio of 1. We report $QE(s_F, h_F)/QE(s_M, h_M)$, i.e., the instance-level ratio of the QE scores assigned to feminine and masculine pairs, in Appendix B.2.

4 Findings

This section presents the results for ambiguous (§4.1), unambiguous with intra-sentential (§4.2.1) and extra-sentential cues (§4.2.2) pairs. Our analysis reveals significant ($p < .05$) gender bias across most evaluated metrics and configurations (see Appendix A.2 for details on statistical testing).

4.1 Gender-Ambiguous Instances

Most QE metrics and language pairs show a significantly higher score for masculine-inflected references. Figure 2 reports the average f/m ratio of QE scores on MT-GenEval test sets (full raw scores are in Figure 5 of Appendix B.1). We found no significant difference between f/m scores only for MetricX (X and XL), en→de. The metric is generally the closest to parity across languages. We hypothesize that MetricX models, unlike Comet, are trained on additional synthetic data that attributes perfect scores to reference translations. This choice likely skews the metric results by making it more sensitive to identifying references than quality differences from minimal edits. Among the best models, Mistral 7B reaches an average ratio of 0.9961. Notably, preference toward the masculine-inflected translations increases with size for both CometKiwi (22 < 23 XL < 23 XXL) and GEMBA models (Mistral 7B < Gemma 2 9B < Llama 3.1 70B). These results hold for GATE (see Figures 6 and 7 in Appendix B.1).

Most QE metrics score gender-neutral translations consistently lower than gendered ones on all languages. Figure 8 in Appendix B.1 reports the ratio between QE scores for gendered and gender-neutral translations on mGeNTE's ambiguous instances. Despite the unknown gender in English, all metrics but MetricX and Mistral 7B prefer gendered translations significantly. Similar to the f/m setup, we observe that the larger a CometKiwi variant is, so is the preference for gendered translations. Findings are consistent across languages, with a larger magnitude on en→es, suggesting that QE metrics may still be unsuitable for assessing gender-fair forms (Sczesny et al., 2016).

4.2 Gender-Unambiguous Instances

Tables 1 and 2 present the total error rate (ER) and error rate ratio ($\Phi(S^F, S^M)$) for the intra and extra-sentential cases across metrics.

4.2.1 Intra-sentential Cues

Most metrics consistently exhibit a higher frequency of errors when the source mentions a feminine entity than a masculine one. As shown in Table 1, the majority of metrics yield values greater than 1 for $\Phi(S^F, S^M)$, indicating increased susceptibility to errors when tasked to select the correctly inflected translation and the referent in the source is feminine. CometKiwi 23 XXL is the only exception, yielding the least biased behavior overall; however, it still exhibits bias in 2 out of 8 languages (full results disaggregated by language are in Table 8 of Appendix B.2).

[6] We count ties as errors since they indicate the system's failure to select the correct gender inflection.

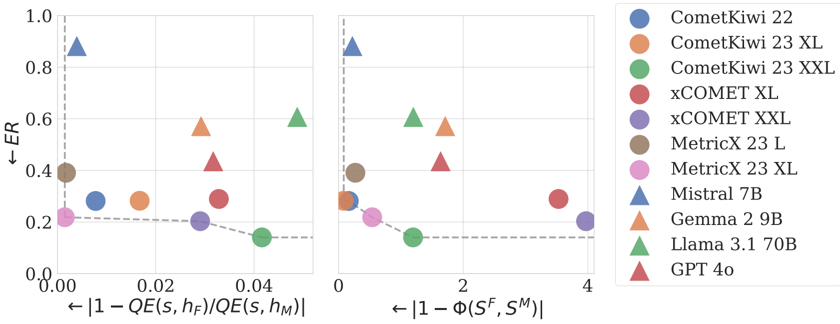

Figure 3: **Total error rate vs. Gaps from parity.** X-axes show gaps from parity for ambiguous (left) and unambiguous (right) instances. Dashed lines indicate the Pareto frontier. Ideally, a metric should achieve both a low error rate and parity, resulting in proximity to the origin (0,0). Neural metrics use the original context.

Metric	ER	Φ	$\mathrm{tr}(c) = \checkmark$ ER	$\mathrm{tr}(c) = \checkmark$ Φ
CometKiwi 22	0.28	0.84	0.19	6.68
CometKiwi 23 XL	0.28	<u>**0.92**</u>	0.17	5.09
CometKiwi 23 XXL	<u>**0.14**</u>	2.20	<u>**0.14**</u>	4.85
xCOMET XL	0.29	4.53	0.24	7.03
xCOMET XXL	0.20	4.97	0.20	5.47
MetricX 23 L	0.39	0.73	0.34	1.53
MetricX 23 XL	0.22	1.54	0.28	1.60
Llama 3.1 70B	0.61	2.20	-	-
Gemma 2 9B	0.57	2.71	-	-
Mistral 7B	0.88	**1.22**	-	-
GPT 4o	**0.44**	2.64	-	-

Table 2: **Total error rate** (ER ↓) **and error rate ratio between groups** ($\Phi(S^F, S^M) \to 1$). MT-GenEval *contextual* set, unambiguous (extra-sentential). $\mathrm{tr}(c) = \checkmark$: context-aware metrics with translated context. Best models per type in bold, best overall, underlined. Mean results across eight languages.

4.2.2 Extra-sentential Cues

Most QE metrics exhibit a significantly higher error rate on feminine sources than masculine sources. When the context is not translated, four out of seven context-aware neural metric variants make at least 1.5 times as many errors on feminine sources as, on average, masculine ones. CometKiwi 23 XL yields the best error rate ratio among neural metrics, followed by CometKiwi 22 and MetricX 23 L.

Metrics that use translated context yield lower overall error rates. However, improved accuracy is primarily driven by better results on masculine sources. Therefore, **translated context exacerbates gender disparities for all QE metrics** (see full results in Table 10 of Appendix B.2). $\Phi(S^F, S^M)$ increases on average 3 times, with the highest increase accounting for CometKiwi 22 (approximately 8 times higher error rate).[7] We hypothesize that this significant increase might be attributed to the translated context, which may induce bias towards masculine referents. These results suggest that only looking into metrics optimizing the overall error rate can be misleading and result in higher disparities.

Regarding GEMBA metrics, only Mistral 7B shows a balanced error rate between genders. On the contrary, Llama 3.1, Gemma 2, and GPT-4o make approximately two to three times as many errors on feminine sources as on masculine sources. These findings indicate a strong preference for masculine-inflected translations, even when gender cues are available in the preceding context.

However, upon a closer inspection, we noted that **GEMBA QE metrics cannot distinguish gender inflections successfully.** Ties—i.e., an equal DA score for the correct and incorrect translation—are widespread. On average, Mistral 7B assigns the same score to 85%, Gemma 2 9B to 50%, Llama 3.1 70B to 55%, and GPT-4o to 38% of test samples (full details in Table 14, Appendix B.2). Even with fewer ties, (Gemma 2 and GPT-4o) metrics show a significantly ($p < .05$) higher error rate for feminine sources. See Table 11 of Appendix B.2 for results on individual languages. We explore further GEMBA models' brittleness to surface-level differences in the following section.

4.3 Bias and Surface-Level Brittleness

An unbiased QE metric evaluates masculine and feminine translations equivalently when the source gender is unknown. However, whether such scores derive from an unbiased assessment or failure to detect gender markers remains unresolved, which

[7]We observed similar trends even when using Google Translate for translating c (see Appendix B.2).

could compromise accuracy and usability when gender is known. To investigate this distinction, we analyze the metrics results jointly across ambiguous and unambiguous scenarios.

Figure 3 reports the total error rate when source gender is known—a measure of overall accuracy—vs. gender bias measurements in ambiguous ($\text{QE}(s, h_F)/\text{QE}(s, h_M)$) and unambiguous cases ($\Phi(S^F, S^M)$). For simplicity, we report one minus these values to have the optimal result at 0. The chart reveals that among the metrics identified as least biased metrics in the ambiguous case (§4.1), only MetricX 23 XL can distinguish gender inflections in the unambiguous case (ER $<$ .25). In contrast, Mistral 7B scores many ties (Table 14 in Appendix B.2), demonstrating a low bias in both setups but an astonishingly high error rate (and extreme brittleness to gender-related surface-level differences). More generally, this shortcoming can be attributed to the tendency of GEMBA metrics to assign coarse-grained scores (Zheng et al., 2024; Stureborg et al., 2024)—typically 85, 90, 95, and 100 in our experiments. Based on this result, we discourage using GEMBA-prompted language models if capturing gender-related phenomena is crucial for the use case at hand.

CometKiwi 23 XXL is a better alternative, being the only other Pareto-optimal metric in ambiguous and unambiguous setups with a low overall error rate. Comparing the xCOMET and Kiwi families, we note that they achieve similar gender error rates, but the latter is fairer, on average. This holistic view is critical, as it can inform the selection of metrics for downstream applications where accurate and unbiased evaluations are necessary.

5 Downstream Implications

Previous sections have collected evidence of widespread gender bias in contemporary QE metrics. This section explores the downstream implications of using biased QE metrics in different stages of a classical MT pipeline.

5.1 Quality Filtering

Filtering high-quality parallel data from web crawls is a long-standing challenge in MT research. Evolving from traditional approaches using heuristics (Resnik and Smith, 2003) or noise detection techniques (Taghipour et al., 2011), recent work has adopted QE metrics to filter large corpora (Peter et al., 2023; Alves et al., 2024). Arguably, data retained post filtering influences each subsequent step in the MT pipeline.

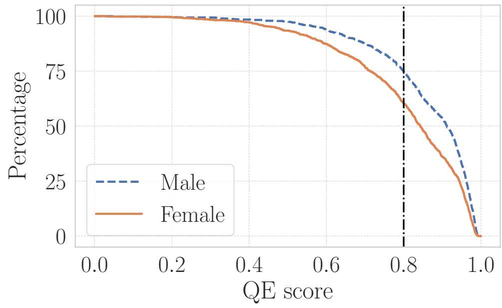

Figure 4: **Share of instances (y-axis, %) scoring at least the value on the x-axis.** Vertical line indicates a typical data filtering threshold. MT-GenEval (ambiguous), En→It, CometKiwi 23 XXL.

Metric	ER_{GT}	$\Phi_{GT}(S^F, S^M)$
CometKiwi 22	0.13	1.34
CometKiwi 23 XL	0.11	1.60
CometKiwi 23 XXL	**0.09**	1.62
xCOMET XL	0.12	1.96
xCOMET XXL	0.10	1.23
MetricX 23 L	0.34	**1.17**
MetricX 23 XL	0.14	1.33

Table 3: **Total error rate** ($\text{ER}_{\text{GT}} \downarrow$) **and error rate ratio between gender groups** ($\Phi_{GT}(S^F, S^M) \to 1$) when assessing *GT translations*. MTGenEval's counterfactual (Unambiguous, Intra-sentential) subset. Best results in bold. Mean results across eight languages.

In §4.1, we show that in the absence of clear gender cues in the source, contemporary QE metrics tend to favor masculine-inflected hypotheses, casting shadows over their use as quality filters. To exemplify the issue, Figure 4 shows a complementary cumulative distribution function of a distribution of CometKiwi 23 XXL scores (Koenecke et al., 2020; Attanasio et al., 2024). The chart illustrates explicitly the number of instances that would be filtered out at a given quality threshold, separately by gender. A threshold of 0.8 would retain 75% of masculine instances but only 63% of feminine ones. This disparity highlights that even when quality is comparable, sentences with feminine forms are more likely to be filtered out, potentially amplifying gender representation imbalances.

5.2 Machine Translation Quality Assessment

Our evaluation in §4 focuses on a controlled setup with minimally contrastive reference translations to isolate the impact of gender-inflected linguistic elements on QE metrics. However, QE met-

Decoding	Metric	Comet-22		Match
		$h_F \uparrow$	$h_M \uparrow$	$\delta_M \sim 0$
Greedy	-	86.73	88.45	-45.67
QAD	CK 22	**87.66**	**89.39**	-46.84
	CK 23 XXL	87.46	89.11	**-43.47**

Table 4: **Quality and gender bias results translating with TowerInstruct-v0.2.** MT-GenEval contextual (ambiguous). Greedy and QAD decoding. Mean scores across six languages. QAD is run with two variants of CometKiwi (CK). Best values are bolded.

rics are primarily used to assess machine-generated translations—which often vary in style and quality (e.g., exhibiting characteristics of *translationese*). We extend our analysis to investigate whether these metrics demonstrate systematic biases when applied to automatic translations. We focus on cases where intra-sentential cues resolve the gender identity of the referent, as in §4.2.1. Specifically, we translate the source-feminine (s_F) and source-masculine (s_M) texts from the counterfactual subset of MTGenEval using Google Translate (GT).[8] To account for quality variations unrelated to gendered forms, we only retain the counterfactual pairs ($s_F, \mathrm{GT_F}$) and ($s_M, \mathrm{GT_M}$) where the translations $\mathrm{GT_F}$ and $\mathrm{GT_M}$ exhibit the correct gender inflection, as categorized by Currey et al. (2022), and are of similar quality as measured by BLEU (Papineni et al., 2002).[9] Our findings reveal that **QE metrics exhibit a significant bias when assessing automatic translations**. When challenged to select the correctly inflected automatic translation, they are more prone to errors on sources with feminine referents than on those with masculine referents (Table 3).[10] These results are in line with those we obtained from human-written references.

5.3 Quality-Aware Decoding

So far, we conceptualized and measured gender bias as an undue preference expressed by quality metrics. However, when used to guide the translation for improved quality, a biased metric can impact and potentially amplify gender bias of MT outputs. To investigate this intuition, we experiment with quality-aware decoding (Fernandes et al., 2022, QAD), a common downstream application of QE metrics. Unlike greedy decoding, which selects the most likely next token at each step, QAD approaches rank N-best translation hypotheses generated by an MT model using automatic metrics.

Setup As the target translation system, we experiment with TowerInstruct-v0.2 (Alves et al., 2024), a state-of-the-art model for MT. Using greedy and QAD, we generate translations of the MT-GenEval contextual (ambiguous) sources for six supported languages: de, es, fr, it, ru, and pt. For QAD, we generate 50 hypotheses using epsilon sampling ($eps = 0.02$) and perform reranking using CometKiwi 22 and CometKiwi 23 XXL. Based on the findings in §4.3, CometKiwi 23 XXL achieves Pareto-optimal performance and CometKiwi 22 exhibits bias in the ambiguous case, allowing us to compare the impact of differently biased metrics.

We evaluate translation quality using reference-based COMET-22 against masculine (h_M) and feminine (h_F) references. Relying on notions linking gender bias and uneven visibility (Savoldi et al., 2021), we measure bias in the MT system as the frequency each gender is represented in the generated hypothesis. Following Currey et al. (2022), we check if specific gender-marked words, uniquely present in either h_M or h_F, are present in the generated hypotheses. We report δ_M, the difference between the number of feminine and masculine reference matches. A δ_M close to zero indicates minimal gender bias as translations are equally matched to female and male references, while a large positive (negative) δ_M suggests that the system is biased towards female (masculine) inflections.

Findings Table 4 presents the translation quality of outputs using both greedy and QAD methods. Using greedy decoding, TowerInstruct-v0.2 produces more masculine-inflected translations of higher quality, as evidenced by absolute higher COMET-22 scores for masculine outputs (h_M) and a large negative δ_M (-45.7). While reranking improves overall translation quality over greedy, as measured by COMET-22, for both gender groups, its effect on gender disparity varies depending on the metric used. Specifically, QAD with the CometKiwi 23 XXL model enhances fairness, achieving the best absolute δ_M values for ambiguous source texts. In contrast, using a biased metric such as CometKiwi 22, further exacerbates gender disparity, as indicated by the more negative δ_M values. Therefore, careful selection and calibration of QE metrics are crucial to ensure that improvements

[8] Prior work has shown that GT is less biased when the context provides sufficient information to disambiguate the gender (Rescigno et al., 2023; Piazzolla et al., 2023).

[9] Details about the filtering process are provided in Appendix A.3. Notably, approximately half of the initial instances are retained, as reported in Table 15 of Appendix A.3.

[10] Individual language results are in Table 13 (App B.2).

in overall translation quality do not inadvertently amplify gender bias.

6 Conclusion

We formalized and measured gender bias in quality estimation metrics commonly used to assess translation quality. We studied translation from English into languages with grammatical gender, isolating the gender lexical phenomenon via minimal contrastive pairs. Our findings revealed concerning patterns. Through extensive experiments across languages, domains, and datasets, we found most QE metrics systematically penalize feminine forms in the target language—both when gender is ambiguous and when it is contextually clear. Similarly, gendered forms are mostly preferred over gender-neutral forms. Moreover, we show that QE bias can affect data quality filtering and quality-aware decoding with MT systems.

We should, indeed, watch the watchers. Undue gendered preferences call for a new focus on how we develop and evaluate QE metrics. Future work should center gender representativeness in training data for QE metrics, including and normalizing gender-neutral instances. Moreover, new scrutiny of QE metrics entails moving beyond simple human alignment to create targeted challenge sets that can expose undue preferences.

Limitations

Our study is characterized by several methodological constraints that merit critical reflection. We limited our investigation to sentence-level MT, which, while representative of prevalent machine translation scenarios, does not cover phenomena potentially arising within more extended contexts, e.g., document-level (Jiang et al., 2022; Raunak et al., 2023; Fernandes et al., 2025) and conversational MT (Agrawal et al., 2024a). The English source sentences examined were derived from naturalistic passages of high quality and relatively limited syntactic variation.

Moreover, our evaluation of neural QE metrics with extra-sentential context, relies on inference-time strategies for integrating contextual information (Vernikos et al., 2022; Agrawal et al., 2024a). We acknowledge that translating context may introduce additional bias depending on the specific translation system used. We leave a more in-depth investigation of this phenomenon—including more resource-intensive solutions, such as targeted fine-tuning or interpretability-driven strategies for selecting few-shot examples to mitigate biases (e.g., gender) in the translation of contextual information (Thakur et al., 2023; Zaranis et al., 2024)—to future work.

Finally, methodologically, we implemented GEMBA-DA in a zero-shot configuration, aligning with contemporary research interested in pushing zero-shot problem-solving (Guo et al., 2025). We defer comprehensive exploration of few-shot and MQM-based prompting strategies (Kocmi and Federmann, 2023a) to future work.

Ethical Considerations

Using gender as a variable requires careful ethical considerations. The corpora we used, initially annotated for gender identity using textual markers like pronouns and titles, present two main concerns. First, gender is a nuanced, evolving attribute, making reliance on textual information alone potentially inaccurate, especially with unverifiable or outdated sources. Second, much of our analysis uses binary gender categories (masculine and feminine). However, we recognize gender as a spectrum and advocate for future studies to account for non-binary and other gender identities. To support gender inclusivity in machine translation, we assessed quality estimation on gender-neutral translations in Italian, Spanish, and German using the mGeNTE corpus.

Acknowledgements

We would like to thank Ben Peters, Dennis Fucci, António Farinhas, Sonal Sannigrahi for their feedback and insightful discussions in the earlier versions of the paper, and Anna Currey for providing valuable help with the MT-GenEval dataset. This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sweta Agrawal, Amin Farajian, Patrick Fernandes, Ricardo Rei, and André F. T. Martins. 2024a. Assessing the role of context in chat translation evaluation: Is context helpful and under what conditions? *Transactions of the Association for Computational Linguistics*, 12:1250–1267.

Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. 2024b. Can automatic metrics assess high-quality translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502, Miami, Florida, USA. Association for Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340, Miami, Florida, USA. Association for Computational Linguistics.

Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2023. Challenging the state-of-the-art machine translation metrics from a linguistic perspective. In *Proceedings of the Eighth Conference on Machine Translation*, pages 713–729, Singapore. Association for Computational Linguistics.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. 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.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems (NeurIPS) 2017*.

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.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Scott Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Carlos Escolano, Graciela Ojeda, Christine Basta, and Marta R. Costa-jussa. 2021. Multi-task learning for improving gender accuracy in neural machine translation. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 12–17, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

Patrick Fernandes, Sweta Agrawal, Emmanouil Zaranis, André FT Martins, and Graham Neubig. 2025. Do llms understand your translations? evaluating paragraph-level mt with question answering. *arXiv preprint arXiv:2504.07583*.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Marina Fomicheva and Lucia Specia. 2019. Taking MT evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Mingqi Gao and Xiaojun Wan. 2022. Social Biases in Automatic Evaluation Metrics for NLG. *arXiv e-prints*, arXiv:2210.08859.

Sarthak Garg, Mozhdeh Gheini, Clara Emmanuel, Tatiana Likhomanenko, Qin Gao, and Matthias Paulik. 2024. Generating gender alternatives in machine translation. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 237–254, Bangkok, Thailand. Association for Computational Linguistics.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8164–8180, Mexico City, Mexico. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.

Manuel Lardelli, Giuseppe Attanasio, and Anne Lauscher. 2024. Building bridges: A dataset for evaluating gender-fair machine translation into German. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7542–7550, Bangkok, Thailand. Association for Computational Linguistics.

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.

Alon Lavie. 2011. Evaluating the output of machine translation systems. In *Proceedings of Machine Translation Summit XIII: Tutorial Abstracts*, Xiamen, China.

Minwoo Lee, Hyukhun Koh, Kang-il Lee, Dongdong Zhang, Minsung Kim, and Kyomin Jung. 2023. Target-agnostic gender-aware contrastive learning for mitigating bias in multilingual machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16825–16839, Singapore. Association for Computational Linguistics.

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

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv e-prints*, arXiv:2207.04672.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There's no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2306.05882*.

Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Singapore. Association for Computational Linguistics.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 300–314, Sheffield, UK. European Association for Machine Translation (EAMT).

Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. 2023. Gender biases in automatic evaluation metrics for image captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8358–8375, Singapore. Association for Computational Linguistics.

Miguel Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024a. Aligning neural machine translation models: Human feedback in training and inference. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).

Miguel Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024b. Aligning neural machine translation models: Human feedback in training and inference. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.

Vikas Raunak, Tom Kocmi, and Matt Post. 2023. Evaluating metrics for document-context evaluation in machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 812–814, Singapore. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T.

Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Argentina Rescigno, Johanna Monti, et al. 2023. Gender bias in machine translation: a statistical evaluation of google translate and deepl for english, italian and german. In *INTERNATIONAL CONFERENCE Human-informed Translation and Interpreting Technology (HiT-IT 2023) Proceedings*, pages 1–11. Incoma Ltd.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139, Bangkok, Thailand. 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, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

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. *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.

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA. Association for Computational Linguistics.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in psychology*, 7:154379.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. 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.

Ian Stewart and Rada Mihalcea. 2024. Whose wife is it anyway? assessing bias against same-gender relationships in machine translation. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 365–375, Bangkok, Thailand. Association for Computational Linguistics.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large Language Models are Inconsistent and Biased Evaluators. *arXiv e-prints*, arXiv:2405.01724.

Shuo Sun, Francisco Guzmán, and Lucia Specia. 2020. Are we estimating or guesstimating translation quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6262–6267, Online. Association for Computational Linguistics.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 340–351, Toronto, Canada. Association for Computational Linguistics.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

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.

Giorgos Vernikos and Andrei Popescu-Belis. 2024. Don't rank, combine! combining machine translation hypotheses using quality estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12087–12105, Bangkok, Thailand. Association for Computational Linguistics.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Emmanouil Zaranis, Nuno M Guerreiro, and Andre Martins. 2024. Analyzing context contributions in LLM-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14899–14924, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand. Association for Computational Linguistics.

A Experimental Details

A.1 Computing DA Scores

COMET. We use the HuggingFace Hub for all COMET metrics and the official implementation from `https://github.com/Unbabel/COMET`. All metrics produce an inclusive score between 0 and 1, indicating the worst and best translation possible.

MetricX 23. We use the official implementation found at `https://github.com/google-research/metricx`. Since these metrics produce error assessments, with a score between 0 (no errors) and 25 (maximum errors), we rescale the output as $QE(\cdot) = 1 - o(\cdot)/\ 25$, where $o(\cdot)$ is the raw output of the metric.

GEMBA. We use the same decoding configuration and post-processing for all the language models prompted with GEMBA-like prompts. Specifically, we use greedy decoding, with a maximum number of generated tokens of 256. Each model's specific chat template is retrieved from the transformers' implementation (Wolf et al., 2020). We use the instruction-tuned variants and a zero-shot prompt for all open-weight models. Non-contextual and context-aware prompts are reported in Table 6. We experimented two variants to include context in prompts, finding no significant difference among those. We finally used the v1 reported in Table 6.

We post-process each output using regexes to capture the recommended score, e.g., "`i would score this translation (\d+(\.\d+)?).`" Please refer to our repository for the full list of pattern strings. If no patterns are matched, we take the first integer number in the response. Since some models produced numbered lists (e.g., "1. [...]"), we filter out every row with a numerical score lower than ten or without a score. As prompts request a score between 0 and 100, we rescale them to the interval 0 to 1 after extraction. Using these filters, we drop less than 1% of the rows on average.
The Hugging Face checkpoint and OpenAI's model IDs are reported in Table 5.

A.2 Statistical Significance

We confirm the validity of our results using statistical significance tests. Particularly for testing the statistical significance of the error rate ratio between the gender groups ($\Phi(S^F, S^M) \to 1$), we use paired bootstrap resampling ($p < .05$). Similarly for the statistical significance of QE ratios, we use the one-sample t-test ($p < .05$).

A.3 Filtering Details for Machine Translation Quality Experiments

As outlined in 5.2, to explore whether QE metrics exhibit gender bias when assessing machine-generated translations, we translate the source-feminine (s_F) and source-masculine (s_M) texts from the counterfactual subset of MTGenEval with Google Translate (GT). Then the counterfactual pairs, $(s_F, \mathrm{GT_F})$ and $(s_M, \mathrm{GT_M})$ are filtered through the following two-step process:

1. **Target gender inflection:** We exclude counterfactual pairs where either of the translations do not exhibit the correct target gender inflection, as categorized by Currey et al. (2022).

2. **Translation quality consistency:** We remove pairs with translations of different quality, as measured by BLEU (Papineni et al., 2002). Specifically, we split translations into 5 quality categories ("Poor","Fair","Good","Very Good" and "Excellent"), following Lavie (2011). We retain only the pairs of instances for which both translations belong to the same quality category. Additionally, within each quality category, we confirm that BLEU does not significantly differ between gender groups using the Wilcoxon test ($p < .05$).

This filtering process ensures that all retained counterfactual pairs are correctly gender-inflected and of equivalent translation quality. Importantly, we retain around half of the initial instances as indicated in Table 15, most of which are of high-quality according to BLEU.

A.4 Hardware specifications

All our experiments were conducted using 2 NVIDIA RTX A6000 GPUs.

A.5 Discussion on artifacts

In our analysis we use the MT-GenEval (Currey et al., 2022), GATE (Rarrick et al., 2023), and mGeNTE (Savoldi et al., 2025) datasets and they can be freely used for research purposes as they are under the CC-BY-SA-3.0, MIT, and CC-BY-4.0 licenses respectively.

B Additional Results

This section reports additional figures and tables to supplement the findings in the main body.

[1]	CometKiwi 22	`Unbabel/wmt22-cometkiwi-da`
[2]	CometKiwi 23 XL	`Unbabel/wmt23-cometkiwi-da-xl`
[3]	CometKiwi 23 XXL	`Unbabel/wmt23-cometkiwi-da-xxl`
[4]	xCOMET XL	`Unbabel/XCOMET-XL`
[5]	xCOMET XXL	`Unbabel/XCOMET-XXL`
[6]	MetricX 23 L	`google/metricx-23-qe-large-v2p0`
[7]	MetricX 23 XL	`google/metricx-23-qe-xl-v2p0`
[8]	Mistral 7B	`mistralai/Mistral-7B-Instruct-v0.2`
[9]	Gemma 2 9B	`google/gemma-2-9b-it`
[10]	Llama 3.1 70B	`meta-llama/Meta-Llama-3.1-70B-Instruct`
[11]	GPT 4o	`gpt-4o-2024-05-13`

Table 5: **Hugging Face Hub ID** for each QE metrics tested.

Format	**Template**

GEMBA	`Score the following translation from {src_lang} to {tgt_lang} on a continuous scale from 0 to 100, where score of zero means “no meaning preserved” and score of one hundred means “perfect meaning and grammar”.`
	`{src_lang} source: “{source}”`
	`{tgt_lang} translation: “{hyphothesis}”`
	`Score:`

GEMBA$_{ctx}$ v1	`Score the following translation from {src_lang} to {tgt_lang} on a continuous scale from 0 to 100` **`given the preceding context`**`, where score of zero means “no meaning preserved” and score of one hundred means “perfect meaning and grammar”.`
	`{src_lang} source: “{context}”`
	`{src_lang} source: “{source}”`
	`{tgt_lang} translation: “{hyphothesis}”`
	`Score:`

GEMBA$_{ctx}$ v2	`Score the following translation from {src_lang} to {tgt_lang} on a continuous scale from 0 to 100, where score of zero means “no meaning preserved” and score of one hundred means “perfect meaning and grammar”.`
	`You can use the preceding context to evaluate the translation of the source.`
	`{src_lang} source: “{context}”`
	`{src_lang} preceding context: “{source}”`
	`{tgt_lang} translation: “{hyphothesis}”`
	`Score:`

Table 6: **Templates for LMs.** We use GEMBA-like (Kocmi and Federmann, 2023b) prompts to evaluate context-free (top) and context-aware (bottom) translation quality. Edits to the original prompt to account for context are in bold. We used v1 for all experiments.

B.1 Results on Gender Ambiguous Instances

In this part, we provide a comprehensive view of results for the ambiguous case. Specifically, Figures 5 and 6 report the raw QE scores when assessing masculine and feminine references, on MT-GenEval and GATE respectively. In addition, Figures 7 and 8 report the ratios of the QE scores on GATE and mGeNTE datasets respectively.

B.2 Results on Gender Unambiguous Instances

Additional results for Intra-sentential Case. In this setup, *gender ambiguity is resolved through contextual cues within the sentence itself.* To provide a more comprehensive analysis, we include additional results for this case. Table 7 presents the complete results, including the total error rate ER, error rate ratio between gender groups $\Phi(S^F, S^M)$, and QE score ratios $\text{QE}(s_F, h_F)/\ \text{QE}(s_M, h_M)$, averaged across languages. Additionally, Tables 8 and 9 provide the per-language error rate ratios between gender groups $\Phi(S^F, S^M)$ and per-language QE score ratios $\text{QE}(s_F, h_F)/\ \text{QE}(s_M, h_M)$, along with the corresponding results of statistical significance tests.

Additional results for Extra-sentential Case. In this setup, *gender ambiguity is resolved by contextual cues from preceding sentences*. As explained in §3.1, we experiment with two variants for neural QE metrics: 1) prepending the context of the source *unmodified* to the candidate hypothesis, and 2) prepending the context *translated* to the candidate hypothesis. For GEMBA metrics we experiment with two versions for including context(see Table 6).

To provide a comprehensive view, we include the complete results. Table 10 presents the complete results, including the total error rate ER, the error rates for sources with masculine referents $\mathrm{ER}(S^M)$ and feminine referents $\mathrm{ER}(S^F)$, and the error rate ratio between gender groups $\Phi(S^F, S^M)$, averaged across languages. Table 11 provides the error rate ratios between gender groups $\Phi(S^F, S^M)$, along with the corresponding results of statistical significance tests for individual languages. In addition, Table 14 summarizes the aggregated tie rates across all languages.

According to our findings, the majority of QE metrics is biased. Additionally, translating the context versus prepending it unmodified to the hypothesis, might improve accuracy but in the trade of bias.

Lastly, as indicated in Figure 9, the translation system being used to translate the context—Google Translate (■) or NLLB (●)—does not affect any trend majorly. Metrics might show marginal improvements in accuracy or gender equality, however the majority of them is biased.

B.3 Results when assessing QE metrics on Automatic Translations

In § 5.2 we explore whether QE metrics demonstrate any systematic biases when applied to automatic translations. As detailed in § 5.2 we focus on the intra-sentential case, where the gender identity of the referent is resolved by the source sentence itself. We conduct our analysis on counterfactual pairs $(s_F, \mathrm{GT_F})$ and $(s_M, \mathrm{GT_M})$, by combining the counterfactual subset of MTGenEval and outputs obtained from Google Translate (see Appendix A.3).

To provide a comprehensive view, we include the complete results on individual language pairs. Specifically, in Table 12, we report the QE score ratios, along with the corresponding statistical tests, for individual languages. In addition, Table 13 presents the per-language error rate ratios $\Phi_{GT}(S^F, S^M)$ between gender groups. As shown, when QE metrics are challenged to select the correct gender inflection for an automatic translation, they demonstrate significantly higher error rates on sources with feminine referents than on their masculine counterparts. A finding inline with our observations when assessing human-written translations.

C AI Assistants

We have used Github Copilot[11] and Claude 3.5 Sonnet during development of our research work.

[11] https://github.com/features/copilot

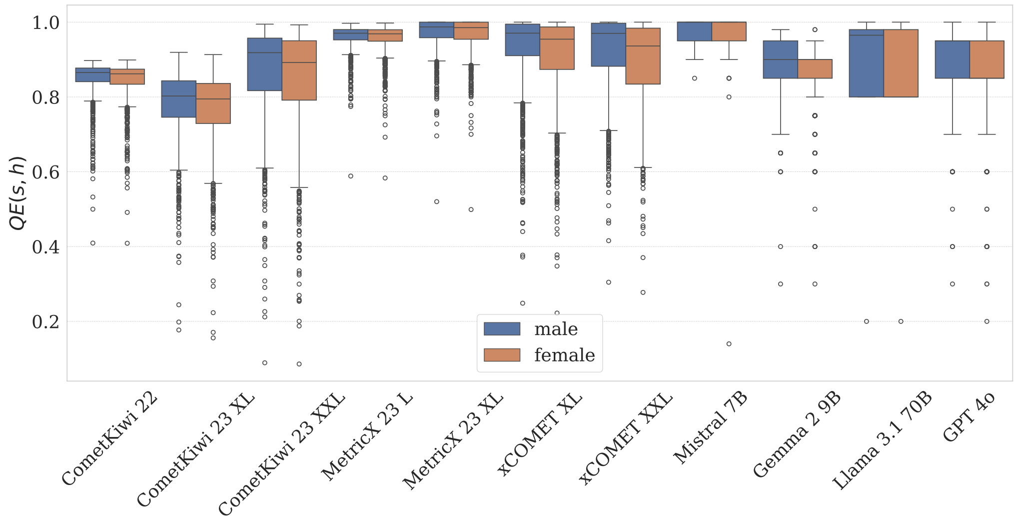

Figure 5: **Raw QE scores for all QE metrics on MT-GenEval.** Ambiguous case. Contextual subset, test instances. Averaged results across languages.

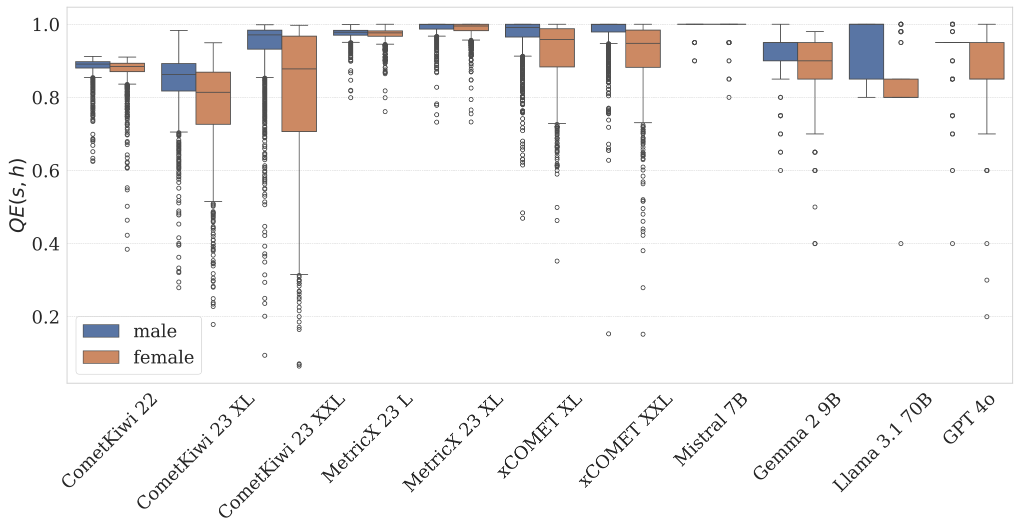

Figure 6: **Raw QE scores for all QE metrics on GATE.** *Ambiguous case.* Averaged results across three languages (fr, es, and it).

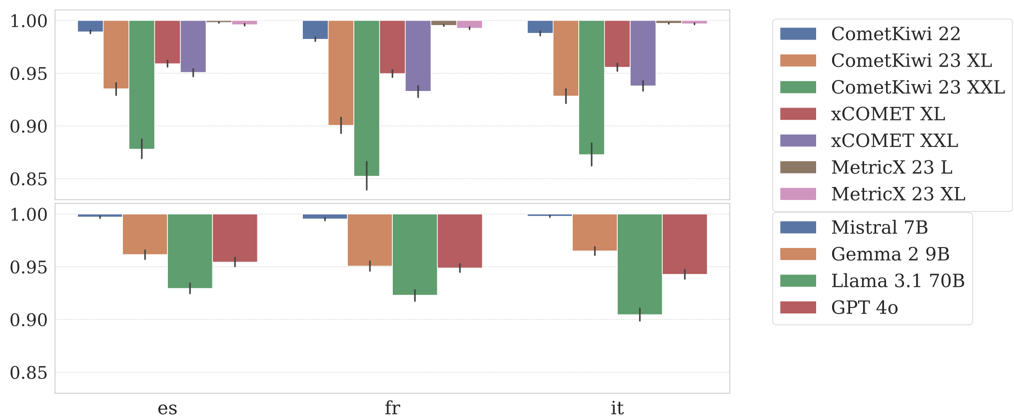

Figure 7: **Ratio** $QE(s, h_F)/\ QE(s, h_M)$ **on GATE.** *Ambiguous case.* Average and 95% confidence interval on the test set by language. Neural QE metrics (top) and GEMBA-prompted language models (bottom).

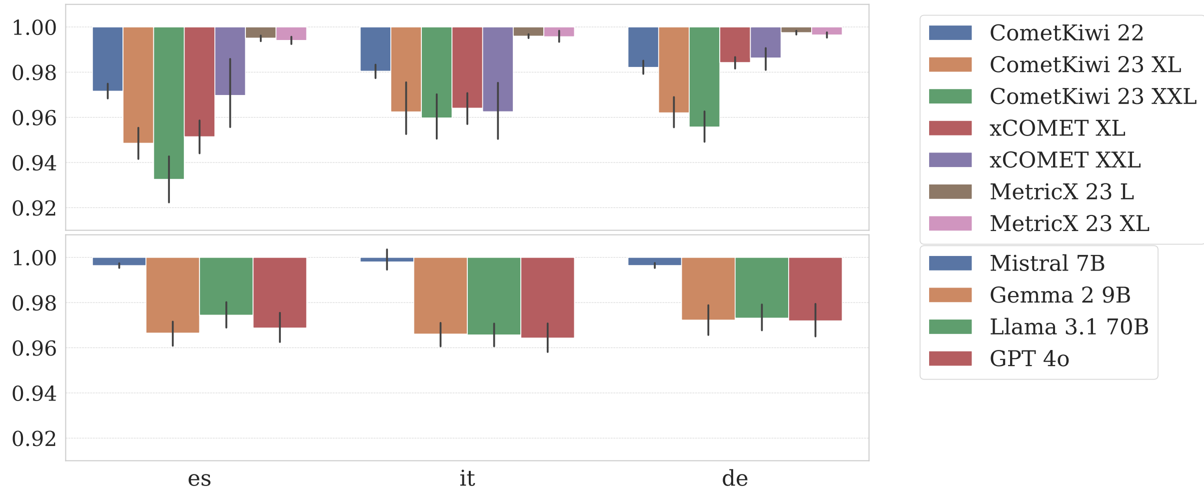

Figure 8: **Ratio** $\mathrm{QE}(s, h_N)/\ \mathrm{QE}(s, h_G)$ **on mGeNTE (Set-N).** *Ambiguous case.* h_N and h_G denote the neutral and gendered translations, respectively. Average and 95% confidence interval on the test set.

Metric	ER	$\Phi(S^F, S^M)$	$\mathrm{QE}(s_F, h_F)/\ \mathrm{QE}(s_M, h_M)$
CometKiwi 22	0.11	1.70	0.997
CometKiwi 23 XL	0.09	1.18	1.001
CometKiwi 23 XXL	**0.07**	**0.87**	1.003
xCOMET XL	0.10	1.81	0.988
xCOMET XXL	0.08	1.32	0.995
MetricX 23 L	0.31	1.25	0.998
MetricX 23 XL	0.12	1.19	**1.000**
Llama 3.1 70B	0.31	1.16	0.995
Gemma 2 9B	0.28	1.36	**1.000**
Mistral 7B	0.74	1.13	**1.000**
GPT 4o	0.16	1.15	1.001

Table 7: **Total error rate** (ER $\downarrow$)**, error rate ratio between gender groups** ($\Phi(S^F, S^M) \rightarrow 1$) **and QE score ratios**. Unambiguous, Intra-sentential case. Counterfactual subset of MT-GenEval. Mean results across eight languages.

Metrics	$\Phi(S^F, S^M) = ER(S^F)/\ ER(S^M)$							
	de	es	fr	it	pt	ru	hi	ar
CometKiwi 22	2.500	1.909	2.333	1.773	0.938	2.500	0.683	1.000
CometKiwi 23 XL	1.143	1.136	0.500	1.000	1.333	2.333	1.297	0.706
CometKiwi 23 XXL	0.286	0.923	0.750	1.036	0.692	2.000	1.000	0.273
xCOMET XL	3.000	1.154	2.200	1.346	2.429	2.200	1.181	1.000
xCOMET XXL	0.857	1.333	1.143	1.321	1.111	2.500	1.426	0.833
MetricX 23 L	1.037	1.582	1.459	1.186	0.839	1.302	0.937	1.660
MetricX 23 XL	1.600	1.083	0.857	0.959	0.550	2.182	1.182	1.130
Mistral 7B	1.190	1.125	1.184	1.194	1.098	1.191	1.045	0.990
Gemma 2 9B	0.775	1.317	1.767	1.443	1.459	1.421	1.248	1.426
Llama 3.1 70B	0.902	1.265	1.113	1.161	1.140	1.185	1.304	1.210
GPT 4o	0.857	1.200	0.939	1.167	1.037	1.368	1.214	1.400

Table 8: **Per language error rate ratio** $\Phi(S^F, S^M)$ **between gender groups** along with the corresponding statistical significance tests. Highlighted values indicate that error rates for sources with feminine referents are significantly higher than for their masculine counterparts. Unambiguous, Intra-sentential case. Counterfactual subset of MT-GenEval.

Metrics	$QE(s_F, h_F)/QE(s_M, h_M)$							
	de	es	fr	it	pt	ru	hi	ar
CometKiwi 22	0.995	0.995	0.995	0.996	1.000	0.997	0.998	1.000
CometKiwi 23 XL	0.997	1.000	0.998	0.996	0.998	1.000	1.006	1.015
CometKiwi 23 XXL	1.005	0.995	1.006	0.991	0.999	1.001	1.006	1.025
xCOMET XL	0.988	0.975	0.997	0.979	0.982	0.991	0.996	0.995
xCOMET XXL	0.997	0.993	0.997	0.989	1.001	0.998	0.989	0.995
MetricX 23 L	0.999	0.998	0.998	0.998	1.000	0.999	1.000	0.992
MetricX 23 XL	0.998	0.998	0.998	1.000	1.001	1.000	1.002	1.000
Llama 3.1 70B	1.002	0.993	0.989	0.999	0.993	0.991	0.995	1.001
Gemma 2 9B	0.997	1.001	1.002	1.004	0.999	0.998	1.003	0.999
Mistral 7B	1.000	1.000	0.999	1.000	0.998	1.000	1.002	1.002
GPT 4o	0.999	1.003	0.998	1.003	1.000	0.993	1.000	1.007

Table 9: **Per language QE score ratios** along with the corresponding statistical significance tests when assessing correctly-inflected feminine and masculine references. Highlighted values indicate that correctly-inflected feminine references receive significantly lower scores compared to their masculine counterparts. Unambiguous, Intra-sentential case. Counterfactual subset of MT-GenEval.

Metric	CometKiwi 22	CometKiwi 22	CometKiwi 23 XL	CometKiwi 23 XL	CometKiwi 23 XXL	CometKiwi 23 XXL	xCOMET XL	xCOMET XL	xCOMET XXL	xCOMET XXL	MetricX 23 L	MetricX 23 L	MetricX 23 XL	MetricX 23 XL	Llama 3.1 70B	Llama 3.1 70B V2	Gemma 2 9B	Gemma 2 9B V2	Mistral 7B	Mistral 7B V2	GPT- 4o
$\mathrm{tr}(c) = \checkmark$		✓		✓		✓		✓		✓		✓		✓							
ER	0.28	0.19	0.28	0.17	**0.14**	**0.14**	0.29	0.24	0.20	0.20	0.39	0.34	0.22	0.28	0.61	0.60	0.57	0.59	0.88	0.86	0.44
$ER(S^M)$	0.32	0.10	0.30	0.09	0.09	0.08	0.11	0.09	**0.07**	0.09	0.46	0.28	0.17	0.24	0.38	0.38	0.31	0.34	0.79	0.75	0.24
$ER(S^F)$	0.25	0.28	0.26	0.24	**0.19**	**0.19**	0.45	0.37	0.32	0.30	0.33	0.39	0.26	0.32	0.81	0.81	0.81	0.82	0.96	0.95	0.61
$\Phi(S^F, S^M)$	0.84	6.68	**0.92**	5.09	2.20	4.85	4.53	7.03	4.97	5.47	0.73	1.53	1.54	1.60	2.20	2.22	2.71	2.54	1.22	1.29	2.64

Table 10: **Total error rate** ER**, error rate per-group** $\mathrm{ER}(S^M)$ **,** $\mathrm{ER}(S^F)$ **and error rate ratio between gender groups** $\Phi(S^F, S^M)$ for all metric versions examined. Unambiguous, Extra-sentential case. Contextual subset of MT-GenEval. $\mathrm{tr}(c) = \checkmark$: context-aware metrics with translated context. Results are averaged across all languages.

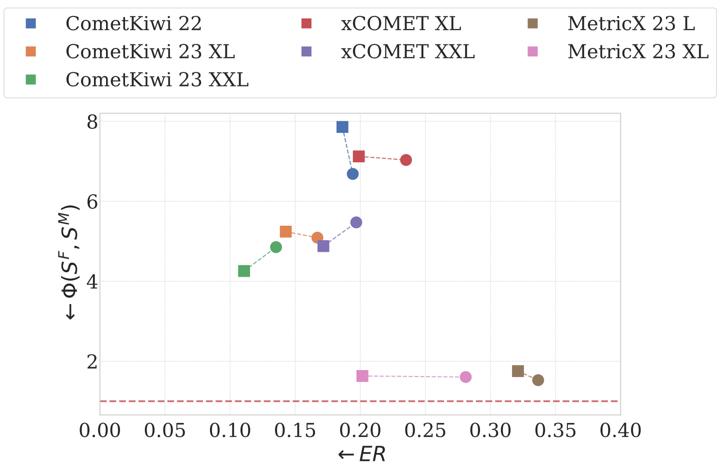

Figure 9: **Total error rate** ER **and error rate ratio** $\Phi(S^F, S^M)$ **between groups for neural QE variants that use translated context.** The translated context is generated by GT (■) or NLLB (●). Unambiguous, Extra-sentential case. Instances of MTGenEval's counterfactual subset . Mean values across eight languages. The red dotted horizontal line indicates parity for Φ.

Metrics	tr(c)	$\Phi(S^F, S^M) = ER(S^F)/ER(S^M)$							
		de	es	fr	it	pt	ru	hi	ar
CometKiwi 22		0.691	0.853	0.895	1.117	1.068	0.724	0.638	0.704
CometKiwi 22	✓	6.636	9.068	8.990	8.278	7.364	10.820	1.506	0.807
CometKiwi 23 XL		1.066	1.012	1.133	1.049	1.248	0.744	0.490	0.595
CometKiwi 23 XL	✓	4.771	4.610	4.362	5.371	6.102	10.396	4.371	0.736
CometKiwi 23 XXL		2.675	2.597	2.221	2.161	2.627	2.027	2.092	1.162
CometKiwi 23 XXL	✓	3.487	3.656	3.508	6.821	6.496	9.941	4.168	0.746
xCOMET XL		5.887	4.686	4.161	4.859	4.308	6.104	4.611	1.622
xCOMET XL	✓	10.291	4.668	10.043	7.486	8.606	7.320	6.547	1.300
xCOMET XXL		4.461	6.011	5.759	6.337	4.848	3.921	5.675	2.760
xCOMET XXL	✓	3.849	5.810	8.804	6.020	5.934	6.153	6.265	0.945
MetricX 23 L		0.559	0.681	0.650	0.821	0.837	0.864	0.823	0.607
MetricX 23 L	✓	1.118	1.476	1.649	1.826	1.050	2.776	1.277	1.038
MetricX 23 XL		1.423	1.423	1.705	1.781	1.051	1.916	1.214	1.802
MetricX 23 XL	✓	1.395	1.353	2.166	2.059	1.300	1.881	1.726	0.955
Mistral 7B		1.494	1.220	1.296	1.285	1.222	1.208	1.018	1.030
Mistral 7B V2		1.665	1.317	1.357	1.324	1.311	1.235	1.035	1.094
Gemma 2 9B		3.129	3.034	2.869	2.654	3.458	2.271	2.127	2.148
Gemma 2 9B V2		2.972	2.935	3.029	2.508	2.861	2.286	1.959	1.806
Llama 3.1 70B		2.126	2.597	2.559	2.397	2.387	1.844	2.238	1.447
Llama 3.1 70B V2		2.195	2.791	2.520	2.279	2.386	1.855	2.253	1.507
GPT 4o		2.705	3.016	2.444	2.811	2.784	2.556	2.202	2.580

Table 11: **Error rate ratio $\Phi(S^F, S^M)$ between gender groups** along with statistical significance tests on individual languages. Highlighted values indicate that error rates for sources with feminine referents are significantly higher than for their masculine counterparts. Unambiguous, Extra-sentential case. Contextual subset of MT-GenEval. $\mathrm{tr}(c) = \checkmark$: context-aware metrics with translated context.

Metrics	$QE_{GT}(s_F, h_F)/QE_{GT}(s_M, h_M)$							
	de	es	fr	it	pt	ru	hi	ar
CometKiwi 22	0.998	0.997	0.996	0.996	0.998	0.996	0.997	0.997
CometKiwi 23 XL	0.999	0.998	0.996	1.011	1.004	0.990	1.006	0.993
CometKiwi 23 XXL	1.003	0.996	1.024	1.000	1.013	0.994	0.996	0.996
xCOMET XL	0.993	0.985	0.985	0.994	0.993	0.983	0.981	1.003
xCOMET XXL	0.999	0.995	0.994	0.999	0.999	0.991	1.005	0.996
MetricX 23 L	1.000	1.000	0.998	0.998	0.999	0.999	1.000	0.992
MetricX 23 XL	1.002	1.001	1.000	0.999	0.997	1.004	1.003	1.001

Table 12: **Per language QE score ratios** along with the corresponding statistical significance tests when assessing correctly-inflected feminine and masculine *GT translations*. Highlighted values indicate that correctly-inflected feminine translations receive significantly lower scores compared to their masculine counterparts. Unambiguous, Intra-sentential case. Source instances from counterfactual subset of MTGenEval.

Metrics	$\Phi_{GT}(S^F, S^M) = ER_{GT}(S^F) / ER_{GT}(S^M)$							
	de	es	fr	it	pt	ru	hi	ar
CometKiwi 22	0.571	1.269	2.000	1.292	1.100	2.333	0.526	1.667
CometKiwi 23 XL	1.600	1.143	0.833	1.160	1.000	3.667	1.053	2.333
CometKiwi 23 XXL	0.600	1.125	4.000	1.000	0.667	1.500	1.030	3.000
xCOMET XL	2.000	1.238	4.000	1.304	2.571	1.800	0.958	1.778
xCOMET XXL	0.667	1.286	1.750	1.038	1.429	1.000	1.167	1.500
MetricX 23 L	1.086	1.216	1.115	1.208	0.973	1.290	0.966	1.474
MetricX 23 XL	1.429	1.207	1.286	1.125	1.400	2.143	1.071	1.000

Table 13: **Per language error rate ratios $\Phi_{GT}(S^F, S^M)$ between gender groups,** along with the corresponding statistical significance tests. *Assessing GT translations*. Highlighted values indicate that error rates for sources with feminine referents are significantly higher than for their masculine counterparts. Unambiguous, Intra-sentential case. Source instances from counterfactual subset of MTGenEval.

Metric	**tr(c)**	**Tie rate (%)**
MetricX 23 L	✓	0.04
MetricX 23 XL		0.19
MetricX 23 XL	✓	11.30
xCOMET XL		0.25
xCOMET XL	✓	0.30
xCOMET XXL		1.44
xCOMET XXL	✓	1.05

(a) Neural metrics

Metric	**Tie rate (%)**
Llama 3.1 70B V1	55.45
Llama 3.1 70B V2	55.01
Gemma 2 9B V1	49.65
Gemma 2 9B V2	51.67
Mistral 7B V1	85.18
Mistral 7B V2	82.48
GPT 4o	38.34

(b) GEMBA metrics

Table 14: **Aggregated tie rates** across all languages for all QE Metrics on the contextual subset of MT-GenEval with at least one tie. $\mathrm{tr}(c) = \checkmark$: context-aware metrics with translated context.

Quality	# Stage 1		# Stage 2
	F	M	
Excellent (50+)	662	654	590
Very Good (40-50)	282	271	191
Good (30-40)	218	209	140
Fair (20-30)	194	189	143
Poor (<20)	149	163	114
Total	1505	1486	1178

Table 15: Each LP originally has 300 instances. We retain approximately 50% of the samples.