

Benchmarking Gender Bias in Machine Translation Evaluation Metrics across Occupations

Orfeas Menis Mastromichalakis¹ * Giorgos Filandrianos² Wafaa Mohammed³
Giuseppe Attanasio¹ Chrysoula Zerva²

¹Instituto de Telecomunicações, Lisbon, Portugal

²National Technical University of Athens, Greece

³University of Amsterdam, Netherlands

Abstract

Gender bias remains a persistent concern in machine translation (MT), affecting both generated translations and their automatic evaluation. When a source text leaves a person’s gender unspecified, translations may realize that person using masculine or feminine forms, and both MT systems and evaluation metrics may exhibit systematic preferences between these alternatives despite the source providing no basis for such a distinction. We study this behavior in the WMT 2026 Automated Translation Quality Evaluation Systems Shared Task using an occupation-balanced subset of GAMBIT+¹. We consider seven English-source language pairs, six from the original dataset, targeting Arabic, Czech, Greek, Icelandic, Russian, and Ukrainian, and extend the original resource with German. The subset contains 1,308 masculine/feminine translation pairs per target language, with three examples for each of the 436 ISCO-08 occupational groups. We evaluate shared-task submissions and baselines for score prediction and error annotation, examining the direction, magnitude, and frequency of gender-related differences. We find an overall tendency for masculine translations to receive higher scores, as well as differences per occupation following stereotypical gender representations, although the strength and consistency of this preference vary considerably across evaluators and languages. Our results show that gender bias remains present in MT evaluation, but that capturing its extent requires looking beyond a single aggregate measure to complementary dimensions of evaluator behavior.

1 Introduction

Gender-related preferences can arise both in machine translation outputs and in the metrics used

to assess their quality. When a source text leaves a person’s gender unspecified, masculine and feminine translations may both be valid, yet MT systems may systematically favor one realization over the other (Menis Mastromichalakis et al., 2025b). This is common for gender-ambiguous occupational terms in English, such as *doctor*, *teacher*, or *legislator*, where the source itself provides no evidence for assigning either gender. Gender-neutral or gender-fair formulations may also be possible in many cases; however, the present work focuses specifically on the contrast between masculine and feminine realizations.

The same concern extends to automatic evaluation. If two translations preserve the information expressed in the source and differ primarily in the gender used to realize an otherwise ambiguous case, an evaluation metric should not systematically reward one variant over the other. Previous work has nevertheless shown that MT evaluation metrics can exhibit such preferences (Filandrianos et al., 2025; Zaranis et al., 2025). This is particularly relevant because automatic metrics are widely used to compare MT systems, rank candidate translations, and guide model development, meaning that systematic evaluator preferences can affect how MT quality is ultimately progressing.

GAMBIT+ (Filandrianos et al., 2025) was introduced to study this form of evaluator bias through controlled masculine/feminine translation pairs. Building on GAMBIT (Menis Mastromichalakis et al., 2025b), it organizes gender-ambiguous occupational references according to the 436 four-digit groups of the ISCO-08 classification² and provides paired translations in which the relevant occupation is realized once in masculine and once in feminine form. The original study evaluated

*Corresponding author: menisorfeas@gmail.com.

¹The dataset is available at: <https://huggingface.co/datasets/ailsntua/gambit-plus>.

²<https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/>

33 source–target language pairs, combining three source languages with eleven target languages, and used the resulting challenge set to analyze quality estimation metrics at WMT 2025.

In this work, we revisit this evaluation for the WMT 2026 Automated Translation Quality Evaluation Systems Shared Task (Lavie et al., 2026) using a smaller, occupation-balanced subset of GAMBIT+. The full benchmark contains nearly 290,000 paired instances, making large-scale evaluation increasingly costly as the number and complexity of evaluation systems grows. We therefore retain three examples for each of the 436 ISCO-08 occupational groups, preserving complete occupational coverage while substantially reducing the computational cost of evaluation. We use English exclusively as the source language and consider seven target languages: Arabic, Czech, German, Greek, Icelandic, Russian, and Ukrainian. Six of these target languages are inherited from the original resource, while German is newly added in this work, yielding 1,308 masculine/feminine translation pairs per target language. We use this benchmark to evaluate both shared-task submissions and baselines for score prediction and error annotation.

However, identifying whether an evaluator exhibits gender bias is not fully captured by a single average score difference. A metric may react strongly to the gendered realization of many examples while showing little net preference because differences in opposite directions cancel out. Conversely, relatively small score differences may consistently favor the same gender across many examples. We therefore examine several complementary aspects of evaluator behavior: signed score differences to capture the direction of preference, absolute paired differences to capture sensitivity regardless of direction, and preference frequencies to measure how consistently one variant is favored. For error-annotation systems, we additionally compare how often masculine and feminine variants are predicted to be error-free.

Our contributions are threefold. First, we construct an occupation-balanced English-source subset of GAMBIT+ for the seven language pairs considered in this work, including the German extension. Second, we evaluate gender-related differences across score-prediction submissions and baselines, jointly analyzing their direction, magnitude, and frequency. Third, we extend the analysis to error-annotation systems, examining whether masculine and feminine translations differ in their

likelihood of being labeled error-free. Overall, our results show that gender bias remains present in MT evaluation, but that its patterns are not fully captured by a single measure and require a more nuanced analysis of evaluator behavior.

2 Related Work

Gender bias in machine translation has been extensively studied across languages, model architectures, and evaluation settings (Savoldi et al., 2021; Vanmassenhove, 2024; Savoldi et al., 2024; Rescigno et al., 2020; Paolucci et al., 2023; Ghosh and Caliskan, 2023; Kostikova et al., 2023; Piazzolla et al., 2023). A particularly relevant strand of this work concerns occupational gender bias, where translation systems may associate professions with stereotypical gender distributions (Menis Mastromichalakis et al., 2025a,b, 2024; Tal et al., 2022; Gorti et al., 2024). Recent work continues to find such effects in modern MT architectures, suggesting that default gender preferences remain an open problem even as model capabilities improve (Manna et al., 2026).

Evaluating these biases is particularly challenging when the source does not provide enough information to determine a person’s gender. Benchmarks such as WinoMT and MT-GenEval focus primarily on cases where contextual evidence identifies the appropriate gender (Stanovsky et al., 2019; Currey et al., 2022), whereas GAMBIT targets occupational references for which masculine and feminine realizations may both be valid (Menis Mastromichalakis et al., 2025b). This distinction has received increasing attention: recent work has explored uncertainty as a way of evaluating model behavior under gender ambiguity (Staliūnaitė et al., 2026), while other studies emphasize preserving ambiguity through gender-neutral or gender-inclusive translation strategies (Dawkins et al., 2025; Savoldi et al., 2025). Our setting is complementary to these approaches, focusing specifically on systematic preferences between masculine and feminine realizations rather than evaluating neutral alternatives.

While most research has focused on biases in MT outputs, less attention has been paid to biases introduced during automatic evaluation. Zaranis et al. (2025) showed that quality-estimation metrics can exhibit systematic gender disparities, and GAMBIT+ extended this analysis to a multilingual, occupation-indexed challenge set (Filandri-

ISCO	1111: Legislators
Source	The Member of Parliament delivered a compelling speech about healthcare reform during yesterday’s session.
Masculine	<u>Der</u> Abgeordnete hielt während der gestrigen Sitzung eine überzeugende Rede über die Gesundheitsreform.
Feminine	<u>Die</u> Abgeordnete hielt während der gestrigen Sitzung eine überzeugende Rede über die Gesundheitsreform.

Table 1: An English source text with the paired German translations introduced in the GAMBITH+ extension. The underlined text marks the gender-dependent difference between the masculine and feminine variants; the remaining text is identical.

anos et al., 2025). More recently, FairQE has explored explicitly mitigating gender bias in quality estimation (Jang et al., 2026). Building on this line of work, we use GAMBITH+ to evaluate WMT 2026 metrics and extend the analysis beyond average score differences, considering complementary dimensions of evaluator preference and sensitivity.

3 The Challenge Set

3.1 GAMBITH+ extension

GAMBITH+ (Filandrianos et al., 2025) provides English texts containing gender-ambiguous occupational references, together with paired target-language translations in which the relevant occupation is realized once in masculine and once in feminine form. The dataset covers the 436 four-digit occupational groups of ISCO-08 and was originally released for multiple source and target languages.

For this year’s evaluation, we focus exclusively on English as the source language and consider seven target languages: Arabic, Czech, German, Greek, Icelandic, Russian, and Ukrainian. Six of these language pairs are inherited from the original GAMBITH+ release, while German is newly added in this work. We follow the same general generation procedure described in GAMBITH+ for constructing paired masculine and feminine translations, using LongCat-2.0³ for the German data. As in the original benchmark, the two target variants are intended to preserve the same source information while differing in the gender realization of the ambiguous occupational reference and any grammatical elements that depend on it. Table 1 shows an example.

³<https://longcat.ai/>

3.2 Occupation-balanced subset

To make the benchmark more practical for evaluating a larger and increasingly computationally demanding set of metrics, we construct a smaller occupation-balanced subset of GAMBITH+. We sample three English source examples for each of the 436 ISCO-08 occupational groups, yielding 1,308 source instances per language pair. This preserves complete occupational coverage while substantially reducing the size of the benchmark. Each source instance is still associated with two target translations: one masculine and one feminine. The resulting benchmark therefore contains seven English–target language pairs, with 1,308 paired translations per target language.

3.3 Evaluated systems

We evaluate the WMT 2026 Automated Translation Quality Evaluation Shared Task systems available for these seven language pairs, including both participant submissions and shared-task baselines. The analysis covers two types of evaluator output: numerical quality scores and error-span annotations.

For score prediction, each evaluator assigns a numerical quality score to the masculine and feminine translation of every source example. For error annotation, systems instead identify translation errors through spans and omission flags. These two settings provide complementary views of evaluator behavior: the former allows us to measure differences in assigned quality scores, while the latter allows us to examine whether masculine and feminine variants are treated differently in terms of predicted errors.

All analyses are performed on the complete set of 1,308 masculine/feminine pairs available for each evaluated language–system combination. We additionally verify the consistency of segment identifiers and pair alignment before computing the results. Appendix A reports the detailed system coverage and score ranges. Since the benchmark is stratified by occupation, the analysis below treats occupation as the primary grouping unit when assessing the stability and consistency of gender-related score differences.

4 Analysis Protocol

4.1 Direction, magnitude and frequency

Let $q_{k\ell i}^M$ and $q_{k\ell i}^F$ be the scores for metric k , target ℓ and sample i , for the masculine and feminine

translations respectively. We define:

$$d_{kli} = q_{kli}^M - q_{kli}^F, \quad (1)$$

$$S_{kl} = \frac{1}{n_\ell} \sum_i d_{kli}, \quad (2)$$

$$A_{kl} = \frac{1}{n_\ell} \sum_i |d_{kli}|. \quad (3)$$

Positive S values indicate higher masculine scores, while A measures sensitivity regardless of direction. In particular, $A \neq |S|$ in general: opposing differences cancel in S but not in A .

For comparability with the range-based analysis of GAMBIT+, we also report

$$S_{kl}^\% = 100S_{kl}/R_k, \quad A_{kl}^\% = 100A_{kl}/R_k, \quad (4)$$

where R_k is the maximum minus minimum *individual* score over both variants and all available target languages for that submission. These percentages express the differences as a fraction of the observed range. They remain sensitive to outliers and coverage.

We additionally calculate masculine-win, feminine-win and tie rates, with all pairs considered in the denominator. The preference balance then corresponds to $100[\Pr(d > 0) - \Pr(d < 0)]$ percentage points. For $A > 0$, the cancellation ratio $1 - |S|/A$ quantifies the discrepancy between net and absolute effects.

4.2 Aggregation, confidence intervals, and significance testing

Each target language contains three examples for each of the 436 ISCO-08 occupational groups. Since every occupation is represented equally, averaging over all 1,308 translation pairs is equivalent to first averaging within each occupation and then averaging over occupations. For multilingual summaries, we compute macro-averages across languages, giving each language equal weight. We report averages over all available languages for each system, and use a fixed six-language panel (AR, CS, DE, IS, RU, and UK) when comparing the 22 systems with common coverage.

To assess the stability of the aggregate effects across occupations, we use an occupation-block bootstrap. We resample the 436 occupations with replacement, retaining the three paired examples belonging to each selected occupation, and recompute the signed and absolute differences over 5,000 bootstrap replicates. For multilingual summaries,

all observations associated with the same occupation across the included languages remain in the same block. We report percentile 95% intervals while keeping the observed normalization ranges fixed. Because the benchmark already covers all 436 ISCO-08 groups, these intervals should be interpreted as sensitivity to changes in the relative weighting of occupations, rather than as sampling uncertainty over unseen occupations or languages.

For each system–language run, we test whether the mean signed difference across the 436 occupations differs from zero using a two-sided one-sample t -test. We correct the resulting 148 p -values for multiple comparisons using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995). A significant result indicates a consistent directional preference across occupations; the sign of the mean determines whether masculine or feminine variants receive higher scores on average.

5 Score-Prediction Results

5.1 Systematic preference and cancellation

Table 2 reports both normalized summaries for every Task 2 (Quality Score Prediction) system. After Benjamini–Hochberg correction at $q = 0.05$, 131 of the 148 system–language runs show a statistically detectable directional difference. Of these, 126 favor masculine variants on average and five favor feminine variants. The five significant negative runs are VERTICAL_870257 on Arabic, VERTICAL_870554 on Arabic and German, and FLUENCY2-GEMINI35-ESA on German and Ukrainian. Significance should also be read together with magnitude: a consistent but small preference and a large, heterogeneous effect have different implications.

Figure 1 compares the signed and mean absolute differences on the common six-language set. The contrast between the two quantities is particularly clear for FLUENCY2-GEMINI35-ESA: although its mean absolute difference is large ($A = 5.7234$, $A^\% = 12.69$), its signed mean is close to zero ($S = -0.0996$, $S^\% = -0.22$). This indicates that substantial gender-related score changes occur in both directions and largely cancel in the signed average. Qwen 3.6, by comparison, has $S^\% = 4.06$ and $A^\% = 7.34$, while xCOMET XXL reference-free has $S^\% = 5.75$ and $A^\% = 7.01$, indicating a more consistently positive preference relative to the overall magnitude of the differences.

COMETKIWI22 (Rei et al., 2022) illustrates a

Submission	AR	CS	DE	EL	IS	RU	UK	Macro-6
AEGIS	+0.66/1.20	-0.03/1.79	+0.26/0.57	+0.45/0.74	+0.61/1.96	+0.67/1.47	+0.19/1.13	+0.39/1.35
†COMETKiwi22	+0.91/1.38	+0.52/0.94	+0.20/0.57	+0.68/0.82	+1.62/1.95	+1.11/1.37	+0.46/0.95	+0.80/1.19
Cohere CAT+ ensemble	+3.56/5.07	+2.68/4.63	+1.08/4.10	+2.19/3.48	+2.59/5.07	+5.13/6.65	+2.12/4.51	+2.86/5.01
FACET_869543	+2.21/3.99	+2.18/4.16	+0.38/4.06	+2.32/4.72	+1.77/4.11	+3.09/5.04	+1.37/3.81	+1.83/4.19
FACET_869546	+2.26/4.29	+2.17/4.33	+0.43/4.24	+2.39/4.81	+1.81/4.17	+3.16/5.22	+1.40/3.86	+1.87/4.35
†Gemini-3.6-Flash	+3.02/5.25	+1.57/3.59	+0.08/3.01	–	+2.19/4.06	+3.24/4.39	+1.17/3.14	+1.88/3.91
†Gemma 4	+8.91/14.17	+3.68/5.24	+0.84/3.46	–	+4.62/6.33	+5.70/6.71	+3.71/5.35	+4.58/6.88
†Gemma 4 RB	+8.93/14.25	+3.61/5.22	+0.74/3.27	–	+4.62/6.33	+5.48/6.65	+3.71/5.35	+4.52/6.85
Lexicala (Majority-Vote)	+11.47/14.58	–	–	–	–	–	–	–
Lexicala (STAPLE)	+12.11/15.72	–	–	–	–	–	–	–
MQM-LLM	+3.51/5.04	+2.82/4.68	+2.08/4.46	+2.70/4.18	+4.37/6.56	+4.93/6.08	+4.11/5.67	+3.64/5.41
MQM-LLM CA	+3.64/5.22	+3.07/5.30	+2.08/5.17	+3.29/4.85	+4.85/7.20	+5.24/6.74	+4.48/6.36	+3.89/6.00
†Qwen 3.6	+3.89/7.59	+3.98/6.92	+2.09/6.15	–	+4.05/8.50	+5.77/7.84	+4.58/7.03	+4.06/7.34
†Qwen 3.6 RB	+3.80/7.52	+3.95/6.92	+1.72/5.47	–	+4.05/8.50	+5.82/7.95	+4.58/7.03	+3.99/7.23
Vertical_870257	-0.24/2.67	+0.04/1.79	+0.07/1.31	+1.23/1.68	+0.46/1.54	+0.63/1.39	+0.95/1.82	+0.32/1.75
Vertical_870554	-0.46/2.41	-0.11/2.64	-0.34/2.51	+2.34/3.19	+0.33/2.49	+1.06/2.55	+0.82/1.76	+0.22/2.39
bytepop_pro10	+3.32/4.56	+0.96/2.33	-0.08/1.79	+3.76/5.59	+4.44/5.85	+2.94/3.42	+2.54/4.26	+2.35/3.70
bytepop_pro9	+3.09/4.16	+0.96/2.33	-0.08/1.79	+3.76/5.59	+3.71/5.25	+2.94/3.42	+2.54/4.26	+2.19/3.54
cuni-v14-regression	–	+8.36/11.58	–	–	–	–	–	–
fluency2-gemini35-esa	-0.61/14.04	-0.92/14.25	-2.59/8.93	+1.20/16.30	+1.35/13.17	+3.40/12.11	-1.97/13.63	-0.22/12.69
gemba-poly	+0.84/1.49	+0.17/0.93	-0.07/1.35	+0.12/1.24	-0.00/1.05	+1.01/2.04	+0.10/1.05	+0.34/1.32
†xCOMET XL RB	+2.60/4.77	+2.11/3.86	+0.63/1.51	–	+3.98/4.95	+3.59/4.37	+2.03/3.30	+2.49/3.79
†xCOMET XL RF	+2.60/4.77	+2.11/3.86	+0.63/1.51	–	+3.98/4.95	+3.59/4.37	+2.03/3.30	+2.49/3.79
†xCOMET XXL RB	+4.93/6.60	+5.00/6.56	+0.95/1.93	–	+6.71/7.70	+8.97/9.94	+7.94/9.32	+5.75/7.01
†xCOMET XXL RF	+4.93/6.60	+5.00/6.56	+0.95/1.93	–	+6.71/7.70	+8.97/9.94	+7.94/9.32	+5.75/7.01

Table 2: Signed / mean absolute paired differences ($S\%/A\%$), in percent of each evaluator’s observed score range. Positive signed values indicate higher masculine scores. Each language has 1,308 pairs. Macro-6 equally averages AR, CS, DE, IS, RU and UK; – denotes missing coverage, never zero. RB/RF retain returned reference-based/reference-free labels, and CA abbreviates Confidence Aware; inference configurations were not independently verified. Native scales appear in Appendix A. † denotes a WMT 2026 organizer baseline; unmarked systems are participant submissions.

different pattern. Its six-language mean absolute difference is small ($A = 0.00674$, corresponding to 1.19% of its observed range), yet it assigns higher scores to masculine variants in 78.1% of pairs, compared with 20.8% for feminine variants and 1.1% ties. Thus, a small average magnitude does not necessarily imply balanced preference frequencies. Conversely, GEMBA-POLY produces ties for 61.8% of pairs.

5.2 Language and occupation differences

To compare target languages using the same set of evaluators, we average the normalized differences over the 13 systems available for all seven languages. The mean absolute difference, $A\%$, ranges from 3.14 in German to 4.64 in Icelandic. Russian shows the largest average signed difference ($S\% = 2.72$), while German is much closer to zero ($S\% = 0.26$).

The direction of the preference can also vary across evaluators. For German, for example, Qwen 3.6 has a positive signed difference ($S\% = 2.09$), whereas FLUENCY2-GEMINI35-ESA has a negative one ($S\% = -2.59$). Thus, even within the same target language, evaluators do not always favor the same gender. This suggests that the observed differences are more strongly associated with the

evaluator than with the target language.

We also examine differences across occupations. Table 3 reports the five occupations with the largest feminine- and masculine-preferring signed differences, based on the mean normalized signed difference $S\%$ across the seven target languages. To make these comparisons consistent, all language-level values are averaged over the same 13 systems with coverage for all seven languages.

Several of the strongest differences align with familiar occupational gender stereotypes. Midwifery professionals and child care workers receive higher feminine scores on average, whereas plumbers, hunters, and trappers receive higher masculine scores. This pattern is broadly consistent with observations in our previous GAMBIT+ evaluation. However, the association is not uniform across occupations or languages: for example, Fashion and Other Models shows a feminine preference overall despite positive differences in Arabic and Czech. We therefore treat these results as descriptive evidence of occupation-level variation rather than as a direct measure of gender stereotypes.

6 Error-Annotation Results

Task 1 provides a complementary view of gender-related evaluator behavior by allowing us to exam-

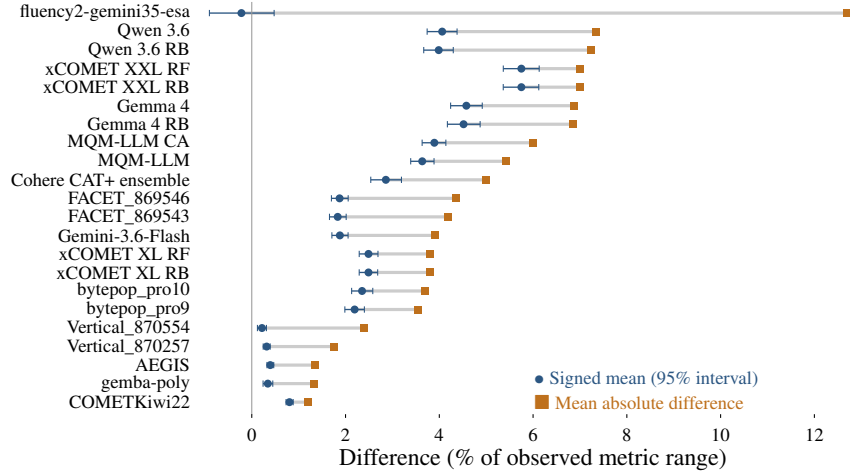


Figure 1: Signed and mean absolute gender-related score differences on the common six-language set across 22 qualifying evaluators. The signed difference captures the direction of preference, while the mean absolute difference captures its magnitude regardless of direction. Whiskers show percentile 95% intervals from 5,000 resamples of the 436 ISCO blocks, keeping paired observations and included languages together. Values closer to zero indicate smaller gender-related differences.

ISCO	Occupation	AR	CS	DE	EL	IS	RU	UK	Mean
2222	Midwifery Professionals	-7.27	-7.43	-4.45	-9.63	-11.79	-5.69	-7.87	-7.73
3222	Midwifery Associate Professionals	-3.09	-4.82	-3.42	-5.70	-10.23	-5.66	-1.85	-4.97
5311	Child Care Workers	-4.24	-0.27	-0.36	-3.03	-5.11	-8.31	-2.57	-3.41
5241	Fashion and Other Models	+4.41	+0.81	-3.25	-4.86	-3.56	-5.34	-10.69	-3.21
5151	Cleaning and Housekeeping Supervisors in Offices, Hotels and Other Establishments	-2.78	-4.49	-2.51	-2.72	-3.79	-3.44	-0.86	-2.94
2636	Religious Professionals	+8.14	+3.28	+4.85	+8.22	+3.34	+4.44	+8.57	+5.84
4213	Pawnbrokers and Money-lenders	+6.77	+3.35	+1.14	+12.32	+5.39	+8.18	+5.62	+6.11
4414	Scribes and Related Workers	+4.09	+2.60	+3.42	+9.90	+5.63	+11.94	+5.53	+6.16
6224	Hunters and Trappers	+6.81	+6.11	+6.52	+9.26	+3.94	+7.05	+7.40	+6.73
7126	Plumbers and Pipe Fitters	+15.34	+3.73	+1.20	+11.83	+4.09	+7.00	+5.55	+6.96

Table 3: Occupation-level extremes across the seven target languages, using the fixed panel of 13 Task 2 systems available for all languages. Each language column reports the mean normalized signed difference $S^{\%}$ across these systems, and the final column gives the macro-average across languages. Negative values indicate higher scores for feminine variants and positive values higher scores for masculine variants.

ine whether masculine and feminine variants differ in how often they are predicted to be error-free. We define a translation as error-free when the system predicts neither an error span nor an omission flag. This closely parallels Task 3 of the shared task, which focuses on the detection of error-free segments. For each system and language, we compute the difference between the masculine and feminine error-free rates, in percentage points; positive values indicate that masculine variants are more often predicted as error-free.

Table 4 reports the results across all available systems and languages. The overall pattern is strongly skewed toward masculine variants: most system-language combinations have positive values, indicating that masculine translations are more of-

ten judged error-free than their feminine counterparts. In several cases the differences are substantial, reaching 27.8 percentage points for GEMMA 4 in Russian and around 30 points for the LEXICALA variants in Arabic.

At the same time, the magnitude of the effect varies considerably across both systems and languages, and the dominant direction is not universal. For example, FLUENCY2-GEMINI35-SPANS favors feminine variants in Czech, German, and Ukrainian, while showing positive differences in the remaining languages. More generally, systems often show similar tendencies across several languages, but different evaluators can behave very differently on the same language. This mirrors the Task 2 results and suggests that the observed

Submission	AR	CS	DE	EL	IS	RU	UK
Cohere CAT+	+8.7	+4.2	+2.2	+4.0	+3.7	+3.1	-0.5
FACET_869532	+4.3	+9.3	+1.3	+4.8	+8.6	+10.1	+20.3
FACET_869533	+4.3	+9.6	+1.2	+4.9	+8.6	+10.2	+20.6
FACET_876357	+4.3	+9.3	+1.3	+4.8	+8.6	+10.1	+20.3
FACET_876365	+4.3	+9.6	+1.2	+4.9	+8.6	+10.2	+20.6
† Gemini-3.6-Flash	+16.4	+7.0	+0.3	–	+7.3	+13.4	+4.5
† Gemma 4	+15.1	+19.1	+3.1	–	+16.1	+27.8	+18.2
† Gemma 4 RB	+15.1	+19.3	+4.1	–	+16.1	+28.4	+18.2
Lexicala (Majority-Vote)	+30.0	–	–	–	–	–	–
Lexicala (Majority-Vote)_876237	+30.0	–	–	–	–	–	–
Lexicala (Majority-Vote)_876371	+30.0	–	–	–	–	–	–
Lexicala (STAPLE)	+30.3	–	–	–	–	–	–
Lexicala (STAPLE)_876238	+30.3	–	–	–	–	–	–
Lexicala (STAPLE)_876370	+30.3	–	–	–	–	–	–
† Qwen 3.6	+7.9	+10.3	+6.3	–	+4.7	+12.1	+9.7
† Qwen 3.6 RB	+8.0	+10.2	+6.0	–	+4.7	+13.5	+9.7
QwenSpan	–	–	+2.1	–	–	–	–
XCOMOffset	–	–	–	–	+1.2	+11.9	+3.3
XCOMOffset_875860	–	–	–	–	+1.2	+11.9	+3.3
XCOMOffset_876378	–	–	–	–	-0.2	+0.0	+0.0
XCOMOffset_876381	–	–	–	–	+1.2	+11.9	+3.3
cuni-cat-v4	–	+21.1	–	–	–	–	–
cuni-v14	–	+21.3	–	–	–	–	–
fluency2-gemini35-spans	+0.8	-0.5	-2.9	+0.4	+2.7	+4.7	-3.5
fluencyscore-multispan-v2	+0.8	-0.5	-2.9	+0.4	+2.7	+4.7	-3.5
fluencyscore-multispan-v2.1	+0.8	-0.5	-2.9	+0.4	+2.7	+4.7	-3.5
† xCOMET XL RF	+2.0	+4.1	+1.7	–	+3.7	+10.7	+5.6
† xCOMET XXL RB	+6.7	+7.8	+2.1	–	+5.4	+17.0	+11.6
† xCOMET XXL RF	+6.7	+7.8	+2.1	–	+5.4	+17.0	+11.6

Table 4: Task 1 differences in predicted error-free rate, in percentage points (masculine minus feminine). Positive values indicate that masculine variants are more often predicted as error-free, while negative values indicate a higher error-free rate for feminine variants. † denotes a baseline; unmarked systems are participant submissions.

gender-related differences are strongly shaped by the evaluator itself rather than a uniform language-level effect.

7 Conclusion

In this work, we revisited gender bias in MT evaluation using an occupation-balanced subset of GAMBIT+, covering seven English-source language pairs and extending the benchmark with German. Across WMT 2026 score-prediction systems, masculine translations are generally favored: 126 of the 131 statistically detectable system–language differences are positive. However, the strength and form of this preference vary substantially across evaluators. Signed differences, absolute differences, and preference frequencies often capture different aspects of the same behavior, while the error-annotation analysis shows a similar overall tendency for masculine variants to be predicted as error-free more often, with important exceptions.

Overall, our results show that gender-related bias remains present in modern MT evaluation systems, but cannot be adequately characterized by a single aggregate measure. The observed effects vary across evaluators, languages, and occupations, with some of the strongest occupation-level differences aligning with familiar gender stereo-

types. We therefore recommend assessing evaluator bias through complementary measures of direction, magnitude, and consistency. The occupation-balanced GAMBIT+ subset introduced here provides a more computationally practical resource for such analyses while preserving complete ISCO-08 occupational coverage.

Limitations

German was generated separately from the inherited GAMBIT+ targets, and the language-specific subsets do not necessarily contain the same English contexts. In addition, each occupation–language combination contains only three examples, so occupation-level differences should be interpreted descriptively rather than as stable estimates.

The analysis also has certain methodological limitations. Range-based normalization is sensitive to extreme scores, and small gender-related differences should not be interpreted as evidence of overall evaluator quality. We also observe only one returned run per system and therefore cannot separate systematic effects from stochastic variation. Finally, our evaluation considers only masculine and feminine variants; gender-neutral or inclusive alternatives are outside the scope of the submitted benchmark. Accordingly, we do not draw causal

conclusions about language effects, occupation-specific stereotypes, or evaluator fairness beyond the pairs and systems studied here.

Acknowledgments

O. Menis Mastromichalakis and G. Attanasio were supported by the project DECOLLAGE (ERC-2022-CoG 101088763), and FCT/MECI through national funds and co-funded EU funds under UID/50008: Instituto de Telecomunicações. G. Filandrianos and C. Zerva were partly supported within the framework of the Pharos AI Factory project, funded by the European High-Performance Computing Joint Undertaking (EuroHPC JU) under Grant Agreement No. 101234269 as part of Horizon Europe and by the Greek Public Investments Program programme. G. Attanasio was additionally supported by the AMALIA project under Measure RE-C05-i08 of the Portuguese national Programa de Recuperação e Resiliência, and by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI). C. Zerva was additionally supported by a Google Research Scholar award. We would also like to thank Manuel Lardelli for the help in reviewing selected German masculine and feminine translations for quality and validity.

References

- Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.
- Anna Currey, Maria Nadejde, Raghavendra Reddy Papagari, 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.
- Hillary Dawkins, Isar Nejadgholi, and Chi-kiu Lo. 2025. Gender-neutral machine translation strategies in practice. In *Proceedings of the 3rd workshop on gender-inclusive translation technologies (GITT 2025)*, pages 74–88.
- George Filandrianos, Orfeas Menis Mastromichalakis, Wafaa Mohammed, Giuseppe Attanasio, and Chrysoula Zerva. 2025. Gambit+: A challenge set for evaluating gender bias in machine translation quality estimation metrics. In *Proceedings of the Tenth Conference on Machine Translation*, pages 314–326.
- 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.
- Atmika Gorti, Aman Chadha, and Manas Gaur. 2024. Unboxing occupational bias: Debiasing llms with us labor data. In *Proceedings of the AAAI Symposium Series*, volume 4, pages 48–55.
- Jinhee Jang, Juhwan Choi, Dongjin Lee, Seunguk Yu, and Youngbin Kim. 2026. Fairqe: Multi-agent framework for mitigating gender bias in translation quality estimation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 37891–37911.
- Aida Kostikova, Joke Daems, and Todor Lazarov. 2023. How adaptive is adaptive machine translation, really? a gender-neutral language use case. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 95–97.
- Alon Lavie, Greg Hanneman, Stefano Perrella, Shuoyang Ding, Eleftherios Avramidis, Lorenzo Proietti, Chi-kiu Lo, Ammon Shurtz, Chrysoula Zerva, Archchana Sindhuja, Vilem Zouhar, Diptesh Kanojia, Frederic Blain, Brian Thompson, Giorgos Filandrianos, Orfeas Menis Mastromichalakis, Tom Kocmi, and Pranav Gupta. 2026. Findings of the wmt26 shared task on automated translation quality evaluation: Compact open models are competitive quality evaluators. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Chiara Manna, Hosein Mohebbi, Afra Alishahi, Frederic Blain, and Eva Vanmassenhove. 2026. [Gender disambiguation in machine translation: Diagnostic evaluation in decoder-only architectures](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 8535–8550, Palma de Mallorca, Spain. ELRA Language Resource Association.
- Orfeas Menis Mastromichalakis, George Filandrianos, Maria Symeonaki, Glykeria Stamatopoulou, Dimitris Parsanoglou, and Giorgos Stamou. 2025a. Gender bias in machine learning: insights from official labour statistics and textual analysis. *Quality & Quantity*, pages 1–35.
- Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Maria Symeonaki, and Giorgos Stamou. 2025b. Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational terms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32221–32237.
- Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Eva Tsouparopoulou, Dimitris Parsanoglou, Maria

- Symeonaki, and Giorgos Stamou. 2024. Gostmt: A knowledge graph for occupation-related gender biases in machine translation. *arXiv preprint arXiv:2409.10989*.
- Angela Balducci Paolucci, Manuel Lardelli, and Dagmar Gromann. 2023. Gender-fair language in translation: A case study. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 13–23.
- 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*.
- 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 Anna Rescigno, Eva Vanmassenhove, Johanna Monti, and Andy Way. 2020. A case study of natural gender phenomena in translation. a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 257–262.
- Beatrice Savoldi, Giuseppe Attanasio, Eleonora Cupin, Eleni Gkovedarou, Janiça Hackenbuchner, Anne Lauscher, Matteo Negri, Andrea Piergentili, Manjinder Thind, and Luisa Bentivogli. 2025. [Mind the inclusivity gap: Multilingual gender-neutral translation evaluation with mGeNTE](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13698–13720, Suzhou, China. Association for Computational Linguistics.
- 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.
- Ieva Staliūnaitė, Julius Cheng, and Andreas Vlachos. 2026. Uncertainty quantification for evaluating gender bias in machine translation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2204–2225.
- 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.
- Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120.
- Eva Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. *Gendered Technology in Translation and Interpreting*, pages 225–252.
- Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and Andre Martins. 2025. [Watching the watchers: Exposing gender disparities in machine translation quality estimation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25261–25284, Vienna, Austria. Association for Computational Linguistics.

A Coverage, Native Scales and Macro Results

Table 5 gives the interval of individual scores observed across each submission’s available languages and the corresponding range R_k . Available-language means use all returned targets for that submission. Six-language means use only AR, CS, DE, IS, RU and UK.

Tables and figures abbreviate reference-based/free as RB/RF, Confidence Aware as CA, and Lexicala-QE-Ensemble as Lexicala. The unsuffixed Gemma 4 and Qwen 3.6 names refer to their returned “Reasoning Baseline” configurations.

Submission	L	Observed interval	R_k	S_{avail}	A_{avail}	S_6	A_6
AEGIS	7	[0.000, 100.000]	100.000	+0.4030	1.2652	+0.3948	1.3522
[†] COMETKiwi22	7	[0.342, 0.907]	0.565	+0.0044	0.0064	+0.0045	0.0067
Cohere CAT+ ensemble	7	[18.333, 100.000]	81.667	+2.2575	3.9098	+2.3355	4.0877
FACET_869543	7	[-42.000, 0.000]	42.000	+0.7989	1.7933	+0.7698	1.7619
FACET_869546	7	[-42.000, 0.000]	42.000	+0.8178	1.8556	+0.7867	1.8280
[†] Gemini-3.6-Flash	6	[0.000, 100.000]	100.000	+1.8793	3.9061	+1.8793	3.9061
[†] Gemma 4	6	[0.000, 100.000]	100.000	+4.5768	6.8781	+4.5768	6.8781
[†] Gemma 4 RB	6	[0.000, 100.000]	100.000	+4.5173	6.8458	+4.5173	6.8458
Lexicala (Majority-Vote)	1	[25.770, 100.000]	74.230	+8.5108	10.8195	–	–
Lexicala (STAPLE)	1	[23.890, 100.000]	76.110	+9.2165	11.9635	–	–
MQM-LLM	7	[6.769, 98.928]	92.159	+3.2272	4.8277	+3.3506	4.9895
MQM-LLM CA	7	[14.518, 102.219]	87.700	+3.3401	5.1159	+3.4152	5.2599
[†] Qwen 3.6	6	[0.000, 100.000]	100.000	+4.0614	7.3392	+4.0614	7.3392
[†] Qwen 3.6 RB	6	[0.000, 100.000]	100.000	+3.9874	7.2333	+3.9874	7.2333
Vertical_870257	7	[0.122, 0.965]	0.843	+0.0038	0.0147	+0.0027	0.0148
Vertical_870554	7	[0.000, 1.000]	1.000	+0.0052	0.0251	+0.0022	0.0239
bytepop_pro10	7	[0.000, 100.000]	100.000	+2.5534	3.9723	+2.3525	3.7023
bytepop_pro9	7	[0.000, 100.000]	100.000	+2.4174	3.8296	+2.1938	3.5359
cuni-v14-regression	1	[43.393, 82.535]	39.141	+3.2708	4.5328	–	–
fluency2-gemini35-esa	7	[54.895, 100.000]	45.105	-0.0081	5.9563	-0.0996	5.7234
gemma-poly	7	[0.000, 100.000]	100.000	+0.3099	1.3081	+0.3421	1.3192
[†] xCOMET XL RB	6	[14.248, 100.000]	85.752	+2.1348	3.2522	+2.1348	3.2522
[†] xCOMET XL RF	6	[14.248, 100.000]	85.752	+2.1348	3.2522	+2.1348	3.2522
[†] xCOMET XXL RB	6	[18.329, 100.000]	81.671	+4.6963	5.7238	+4.6963	5.7238
[†] xCOMET XXL RF	6	[18.329, 100.000]	81.671	+4.6963	5.7238	+4.6963	5.7238

Table 5: Observed score scales and macro means in native units. L is the number of returned target languages. Absolute means are means of pairwise absolute differences, not absolute values of signed means. [†] denotes a baseline; unmarked systems are participant submissions.