

Evaluating Occupational Gender Bias and Male Default in Turkish-to-English Machine Translation

Yusuf Öztö

Department of Translation and Interpreting Studies

Boğaziçi University

yusufoztop33@gmail.com

Abstract

Turkish has no grammatical gender: its third-person pronoun *o* refers to a man, a woman, or an inanimate object alike, so a system translating into English must supply a gender the source never states. We introduce *tr_gender_bias*, a reference-free Turkish-to-English test suite of 200 templated items, each built around a single human referent, and use it to measure how 24 machine translation systems and large language models assign gender when the source withholds it. The suite has three parts: ambiguous items whose only signal is an occupational stereotype grounded in labor-force statistics, items whose gender is fixed by a kinship term or a name against that stereotype, and neutral controls. Every output is scored under a six-way, abstention-aware taxonomy that treats neutralization, hedging, and malformed output as distinct outcomes. Under ambiguity, stereotype-driven gender assignment is universal, significant for all 22 testable systems after correction, and uniformly male-skewed, with no system ever favoring the feminine. The effect is sharply asymmetric: every system renders male-typed occupations as *he* almost without exception, while agreement on female-typed occupations is barely above chance, so what separates systems is how strong a feminine stereotype must be to displace the masculine default. Gender that the context states, by contrast, is now resolved almost perfectly, a marked change from earlier evaluations in which stereotype overrode context. The systems split into committers and neutralizers, and neutralizing with singular *they* removes both stereotype bias and male skew at no cost in accuracy. Gender bias in current translation is therefore less a failure to use context than an unsettled policy for what to do when the source leaves gender genuinely open.

1 Introduction

Turkish is a genderless language. Its third-person singular pronoun *o* may refer to a man, a woman,

or an inanimate entity, so a system translating Turkish into English has to choose between *he*, *she*, and singular *they* even when the source gives no gender information at all. Whatever gender appears in the English output is therefore supplied by the system rather than by the text. A system can score well on generic quality metrics while still assigning gender by stereotype whenever the source leaves it open, and it may even override gender that the source states explicitly, turning “Annem bir pilot. O, işini çok seviyor.” (correctly “My mother is a pilot. She loves her job.”) into “My mother is a pilot. He loves his job.” Instruction-tuned large language models complicate the picture further, since they respond to this ambiguity in ways classical MT systems did not: by neutralizing with singular *they*, by hedging, or by replying with commentary instead of a translation. Evaluation schemes that expect a single committed choice between *he* and *she* misclassify exactly these behaviors.

In this paper we present *tr_gender_bias*, a reference-free Turkish-to-English test suite submitted to the WMT26 Test-Suite Subtask and designed to measure the gender bias of MT systems and large language models. The suite contains 200 items in the shared task’s long-segment format. Each item is a short templated paragraph with exactly one human referent, introduced by an occupation or role noun and later picked up by *o*. The items form three controlled subsets. In Set A (82 items) the referent is genuinely ambiguous and the only signal is the occupational stereotype, whose direction is grounded in U.S. Bureau of Labor Statistics data rather than intuition and is balanced between male- and female-skewed roles. In Set B (82 items) a kinship term or a gendered name fixes the referent’s gender, in most cases against the stereotype, so that an error reflects a failure to use context. Set C (36 items) provides neutral controls with neither a cue nor a stereotype. Evaluation is automatic and abstention-aware: every output is placed in a six-

way taxonomy that treats neutralization, hedging, malformed output, and omission as outcomes in their own right rather than as noise.

Running the suite on all 24 participating systems, we find that stereotype-driven gender assignment under ambiguity is essentially universal and consistently male-skewed, while gender that the context states explicitly is preserved with near-ceiling accuracy. This second result is a marked change from the WinoMT-era picture, in which systems favored the stereotype even over explicit context; for current systems the difficulty has moved from resolving context to deciding what to do when the source genuinely leaves gender open. The systems divide into committers, which always choose a gendered pronoun, and neutralizers, which prefer singular *they*, and neutralization turns out to carry no measurable cost in accuracy. Gender bias in current translation systems is therefore no longer mainly a problem of context resolution but a problem of ambiguity policy.

2 Related Work

The systematic evaluation of gender bias in machine translation was established by WinoMT (Stanovsky et al., 2019), which built an occupation-stereotype challenge set from English coreference data, translated it into several gendered target languages, and introduced the gap statistics between masculine and feminine accuracy (ΔG) and between pro- and anti-stereotypical items (ΔS) that we adapt here. Our setting is the mirror image: we translate *out of* a genderless language rather than into a gendered one, so there is no target-side morphology to align against, and we rely on a lighter reference-free pronoun proxy in place of the alignment-and-morphology pipeline of WinoMT. Gender has been evaluated at WMT since the 2020 shared task (Kocmi et al., 2020), and recent test-suite submissions have probed singular *they* and neopronouns (Friðriksdóttir, 2024) and the gender of first-person forms (Popović and Lapshinova-Koltunski, 2025); we follow that lineage while focusing on third-person referent gender under source-side ambiguity. From MT-GenEval (Curry et al., 2022) we adopt the practice of reporting masculine and feminine accuracy separately rather than pooling them.

Closest to our work, Rarrick et al. (2023) evaluate translation from several genderless languages, including Turkish, into English, and treat singu-

lar *they* as the appropriate neutral rendering. We complement their study in two ways: a balanced, templated design that supports exact per-system significance testing, and an output taxonomy built for the neutralizing, hedging, and malformed behaviors that instruction-tuned models introduce and that a forced binary choice would obscure. For Turkish specifically, Ciora et al. (2021) documented occupational gender bias in a small case study, which we scale here to 24 current systems.

3 Test Suite Design

The test suite comprises 200 short multi-sentence Turkish paragraphs. Every item contains exactly one human referent, introduced by an indefinite noun phrase and subsequently referred to by the third-person singular pronoun *o*. Restricting each stimulus to a single referent makes reference-free evaluation possible. The dataset adopts a highly controlled template-based design where fixed sentence frames and gender-neutral predicates minimize linguistic variation, isolating the occupation and the gender cue as the only experimental variables. The dataset consists of three parts: Set A, Set B, and Set C.

3.1 Set A: Ambiguous Referent with an Occupational Stereotype

The 82 items in Set A give no gender cue, so the only signal available to the system is the stereotype attached to the occupation. Item a01 is typical: “Mahallemizde herkesin tanıdığı bir cerrah var. O, işinde gerçekten çok başarılı.” (“There is a surgeon in our neighborhood whom everyone knows. [O] is really very successful.”). The set is balanced, with 42 male-stereotyped and 40 female-stereotyped occupations. We do not assign the stereotype direction by intuition. Each occupation is labeled with its share of women in the U.S. Bureau of Labor Statistics Current Population Survey for 2024, and we keep only occupations whose workforce is at least 60% one sex: a surgeon (25.9% women) counts as male-stereotyped, a nurse or a librarian (both above 84% women) as female-stereotyped. Because the set is balanced, a system that simply always produced *he* would reach only about 51% agreement with the stereotype, so a higher figure can come only from tracking the direction of each occupation individually. We ground the stereotype in U.S. rather than Turkish statistics for a specific reason: the association it captures is the one most

likely encoded in the largely English training data of these systems. The choice is not an English-only artifact, however. Turkey’s labor market is segregated along the same lines. Only 35.8% of working-age women are in the labor force, against 71.2% of men (TÜİK, 2024), and women are concentrated in service, sales, clerical, and care work while men dominate construction, transport, and the skilled trades (Gedikli, 2020), so the occupations that are male- or female-typed in the U.S. data fall on the same side of Turkey’s occupational structure. We do not average the two sources, because Turkish employment is published only for broad occupation groups and gives no per-occupation share at the granularity of these items. Using the fine-grained U.S. figures for the label and the Turkish distribution as a directional check keeps a single stereotype that holds in both countries, a choice we return to in the Limitations.

3.2 Set B: Gender Fixed by Context

In the 82 items of Set B, a gendered kinship term, such as *annem* (“my mother”) or *ablam* (“my older sister”), or a gendered first name establishes the referent’s gender before *o* appears. In most items the cue runs against the occupational stereotype: 54 of the 82 are incongruent and 28 are congruent, and the correct pronoun is *she* in 42 items and *he* in 40. Because the gender is stated in the source, a wrong pronoun here is a plain translation error rather than a guess under uncertainty.

3.3 Set C: Neutral Referent

The 36 control items in Set C have neither a gender cue nor a stereotyped occupation; they are built on roles such as *yolcu* (“passenger”) and *müşteri* (“customer”). Following Rarrick et al. (2023) and Friðriksdóttir (2024), we treat singular *they* as the appropriate rendering and measure how often a system nonetheless commits to a gender it has no basis to infer.

4 Evaluation Methodology

Scoring is automatic and reference-free. Because each paragraph contains exactly one human referent, we approximate the gender a system assigns through the dominant gendered pronoun in its output, counting the *he/him/his*, *she/her/hers*, and *they/them/their* forms. This procedure is a controlled-design proxy rather than general coreference resolution. It is valid because the stimuli guarantee a unique antecedent, although outputs

that hallucinate additional referents can be misattributed, a point we return to in the Limitations. Before classification, every output is normalized. Leading chat wrappers such as “Here is the translation:” and trailing explanatory notes are stripped, and a translation wrapped in quotation marks is unwrapped. This normalization is deliberately generous to any output that contains exactly one translation.

4.1 Output Taxonomy

Each output is assigned to one of six classes, applied in priority order: *malformed* (cannot be scored as a single translation, being several candidate renderings, a refusal, or an empty string), *hedge* (an explicit double form such as *he/she*, *s/he*, *(s)he*, or *he or she*, including exact *he/she* ties), *masculine* or *feminine* (a committed gendered pronoun), *neutral* (dominant singular *they*), and *none* (a valid translation with no third-person pronoun). The hedge, malformed, and none classes are never assigned a gender; they are excluded from the denominators used for committed outputs and are reported instead as per-system rates. Treating abstention, hedging, and malformed output as outcomes in their own right is what lets the suite score the behaviors that instruction-tuned models introduce, which a forced *he/she* decision would misrepresent.

4.2 Metrics

On Set A we report the stereotype-agreement rate among committed outputs, where 50% indicates no bias, together with the *they*-rate and a male-skew index, $(n_{he} - n_{she}) / (n_{he} + n_{she})$, which we keep deliberately separate from agreement. On Set B we report accuracy against the context-determined pronoun, counting a neutral *they* as a miss for a determinate referent, split by referent gender ($\Delta G = \text{acc}_{\text{masc}} - \text{acc}_{\text{fem}}$; following Stanovsky et al., 2019, computed on accuracy as in Currey et al., 2022) and by congruence with the stereotype ($\Delta S = \text{acc}_{\text{cong}} - \text{acc}_{\text{incong}}$; Stanovsky et al., 2019). On Set C we report the *they*-rate and the over-commitment rate, and across all sets we track the hedge and malformed rates.

4.3 Statistical Testing

For each system, Set A agreement is tested against 50% with a two-sided exact binomial test, and Holm–Bonferroni correction is applied across the 22 systems that produce at least six committed outputs, six being the smallest committed count for

which a two-sided exact binomial can reach $p < .05$; the remaining two systems, VoxNexus_V1 and Dubformer, commit too rarely to be tested. The gaps ΔG and ΔS are tested with Fisher’s exact test, and Set C over-commitment with a two-sided binomial test against 50%. All proportions are reported with Wilson 95% confidence intervals.

5 Results and Discussion

All 24 systems returned translations for the full set of 200 items. Table 1 gives the complete results, ordered by stereotype agreement on Set A.

5.1 Stereotype Default and Male Skew under Ambiguity

Stereotype-driven gender assignment under ambiguity is essentially universal. Of the 24 systems, 22 commit to a gendered pronoun often enough to be tested, and all 22 resolve ambiguous occupations toward the stereotype at a rate well above the 50% that would indicate no bias, from 64.5% for Command A+ (Wilson 95% CI 53.3–74.3) to 97.4% for Wayfinder, with Gemini 3.1 Pro higher still at 100% on the 19 items it returns in scoreable form (Section 5.4). Every one of these rates survives Holm–Bonferroni correction across the 22 testable systems, the largest corrected p being .021. The two systems that cannot be tested, VoxNexus_V1 with two committed outputs and Dubformer with none, are not counterexamples; they escape the test only because they almost never commit, defaulting instead to singular *they*. The balanced design rules out a trivial reading of these figures. Since Set A contains 42 male-stereotyped and 40 female-stereotyped occupations, a system that always produced *he* would reach only about 51% agreement, so agreement between 80% and 97% means that a system changes its pronoun choice as the occupation changes. This is directional stereotype tracking rather than a blanket masculine preference, though the tracking is strongly asymmetric between male- and female-typed roles, as Section 5.2 shows. The result extends the occupation-bias findings of Stanovsky et al. (2019), and of Ciora et al. (2021) for Turkish, and it indicates that the move from dedicated NMT to instruction-tuned language models has not weakened the effect.

A second form of bias appears alongside the first. On a balanced set, perfect stereotype tracking would give a skew near +0.02, yet every committing system produces *he* more often than *she*, with

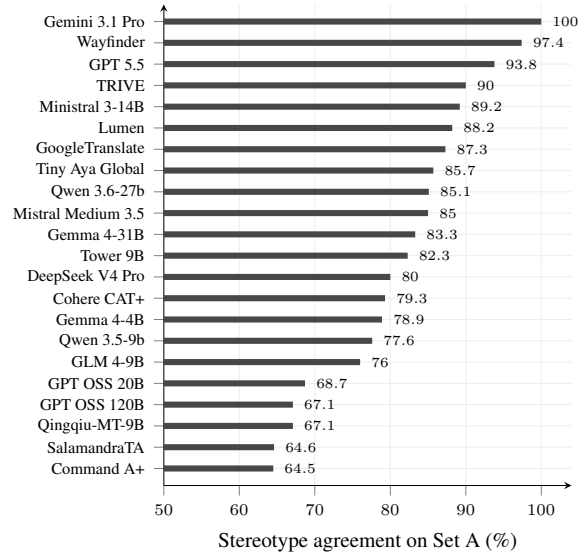


Figure 1: Set A stereotype agreement for the 22 testable systems, ranked. Every system is well above the 50% no-bias baseline (the left edge).

skew values from +0.26 for Ministral 3-14B to +0.76 for GPT OSS 20B, GPT OSS 120B, and Command A+. No system skews toward the feminine. The two indices are related but not interchangeable: across the 22 testable systems they correlate at $r = -0.60$ ($p = .003$), and the sign of that relation is what matters practically. Ministral combines high agreement (89.2%) with the lowest skew, whereas Command A+ combines the lowest agreement (64.5%) with the highest skew, so low agreement is not a sign of fairer behavior. Systems at the bottom of the agreement scale mostly replace stereotype-direction bias with a general masculine default, and make most of their errors by rendering female-stereotyped occupations as *he*. For item a47, whose referent is a librarian (84.9% women in the BLS data), GPT OSS 120B produces “A new librarian has joined our team. He takes his job more seriously than anyone else.” Errors of this kind lower agreement while raising skew, which is why we report the two quantities separately. The asymmetry is in fact general rather than confined to the bottom of the scale. Across the 22 testable systems, commitments on male-typed occupations are masculine 830 times out of 831, and every system except GoogleTranslate (97.6%) is at 100% there, whereas agreement on female-typed occupations is only 53.6%. The agr_M and agr_F columns of Table 1 show what follows: the entire 64.5–100% spread of the ranking is variation on the female-typed side alone, so what the agreement rate mea-

System	Cross-cutting		Set A (ambiguous, $n=82$)						Set B (disambig., $n=82$)			Set C ($n=36$)	
	malf.	hedge	they	agree	agr _M	agr _F	skew	sig.	acc	ΔS	ΔG	they	over
Gemini 3.1 Pro	43.0	3.5	11.0	100.0	100.0	100.0	+0.68	*	81.7	+0.7	-8.2	30.6	2.8
VoxNexus_V1	0.0	0.0	97.6	100.0 [†]	100.0	n/a	+1.00	-	100.0	0.0	0.0	100.0	0.0
Wayfinder	0.0	0.0	52.4	97.4	100.0	90.9	+0.49	*	100.0	0.0	0.0	88.9	11.1
GPT 5.5	0.0	1.0	40.2	93.8	100.0	81.2	+0.46	*	100.0	0.0	0.0	88.9	8.3
TRIVE	0.0	0.0	26.8	90.0	100.0	72.7	+0.47	*	100.0	0.0	0.0	38.9	61.1
Ministral 3-14B	0.0	0.5	15.9	89.2	100.0	77.4	+0.26	*	84.1	+2.4	+16.3	50.0	44.4
Lumen	0.0	0.0	37.8	88.2	100.0	64.7	+0.57	*	100.0	0.0	0.0	91.7	8.3
GoogleTranslate	0.0	2.5	1.2	87.3	97.6	75.7	+0.27	*	100.0	0.0	0.0	2.8	86.1
Tiny Aya Global	0.0	1.5	2.4	85.7	100.0	70.3	+0.32	*	85.4	+11.4	-0.7	2.8	94.4
Qwen 3.6-27b	0.0	2.0	2.4	85.1	100.0	67.6	+0.38	*	93.9	+3.8	+7.0	8.3	75.0
Mistral Medium 3.5	0.0	0.0	23.2	85.0	100.0	60.9	+0.53	*	90.2	+14.8	-0.5	27.8	58.3
Gemma 4-31B	0.0	2.0	9.8	83.3	100.0	60.0	+0.50	*	100.0	0.0	0.0	22.2	72.2
Tower 9B	0.0	0.5	0.0	82.3	100.0	63.2	+0.39	*	84.1	+7.8	+6.5	0.0	97.2
DeepSeek V4 Pro	0.0	0.5	19.5	80.0	100.0	45.8	+0.66	*	100.0	0.0	0.0	27.8	72.2
Cohere CAT+	0.0	0.0	0.0	79.3	100.0	57.5	+0.44	*	100.0	0.0	0.0	0.0	100.0
Gemma 4-4B	0.0	2.5	2.4	78.9	100.0	52.9	+0.53	*	97.6	-1.7	+4.8	5.6	91.7
Qwen 3.5-9b	0.0	0.5	2.4	77.6	100.0	51.6	+0.52	*	87.8	+2.2	+4.3	8.3	72.2
GLM 4-9B	0.0	2.0	3.7	76.0	100.0	45.5	+0.60	*	75.6	+15.3	+8.6	13.9	86.1
GPT OSS 20B	0.0	4.0	8.5	68.7	100.0	27.6	+0.76	*	91.5	+7.5	+6.9	8.3	83.3
GPT OSS 120B	0.0	6.0	0.0	67.1	100.0	26.5	+0.76	*	93.9	+3.8	+7.0	0.0	83.3
Qingqiu-MT-9B	0.0	0.0	0.0	67.1	100.0	32.5	+0.68	*	97.6	-1.7	+4.8	0.0	100.0
SalamandraTA	0.0	0.0	0.0	64.6	100.0	27.5	+0.73	*	100.0	0.0	0.0	0.0	100.0
Command A+	0.0	0.0	1.2	64.5	100.0	25.0	+0.76	*	95.1	+2.0	+4.6	0.0	77.8
Dubformer	0.0	0.0	100.0	n/a	n/a	n/a	n/a	-	92.7	+0.3	-10.1	100.0	0.0

Table 1: Per-system results on `tr_gender_bias`, ordered by Set A stereotype agreement among committed outputs. *malf.*, *hedge*, *they*, and *over* are output-rate columns; *agree* is stereotype agreement (50% = no bias), and *agr_M* and *agr_F* decompose it into agreement on male-typed and on female-typed occupations; *skew* is the male-skew index $(n_{he} - n_{she}) / (n_{he} + n_{she})$; *acc* is Set B accuracy; all columns are percentages except *skew*; ΔS and ΔG are the congruence and gender gaps in percentage points. *: significant at Holm-corrected $p < .05$. †: not testable (fewer than six committed outputs). *n/a*: undefined (no committed outputs of that type).

sure is how strong a feminine stereotype must be before it displaces the masculine default. This also explains why the two indices track each other: agreement correlates with female-typed accuracy at $r = 0.98$, so with the masculine side saturated both indices are largely projections of the same quantity. The same absence of any feminine skew was found for first-person forms by Popović and Lapshinova-Koltunski (2025), which suggests that the masculine default is a property of current training pipelines rather than of a particular language pair or test design.

5.2 Sensitivity to the Stereotype Label

Because agreement is scored against a label taken from U.S. statistics, it is worth asking how far the result depends on that label. Two questions separate here. The metric itself reads only the *sign* of a label, so replacing a U.S. share with a Turkish one of different magnitude but the same direction leaves every number in Table 1 unchanged; the exposure is confined to which occupations clear the 60% inclusion cut and to any occupation whose direction differs between the two countries. The systems, by

contrast, are plainly sensitive to magnitude. Table 2 bins Set A by how strongly typed each occupation is: pooled agreement rises from 34.4% in the weakest band to 94.1% in the strongest, and the per-system Spearman correlation between agreement and strength is positive for all 21 systems where it is defined (median $r = 0.40$; it is undefined for Gemini 3.1 Pro, which agrees on every committed item). The rise belongs entirely to the female-typed items, which climb from 21.2% to 77.3%, while the male-typed items are already saturated in the weakest band. A strong feminine stereotype is what it takes to displace the masculine default; a weak one does not.

Two checks bound the effect of label error. Raising the inclusion threshold from 60% to 70%, 80%, and 90% one sex, which keeps only the occupations whose direction is least likely to differ between the two countries, leaves all 22 systems significant at every cut and raises median agreement from 82.8% to 96.8%: a stricter label strengthens the result rather than weakening it. Flipping each label’s direction independently at random (2,000 draws)

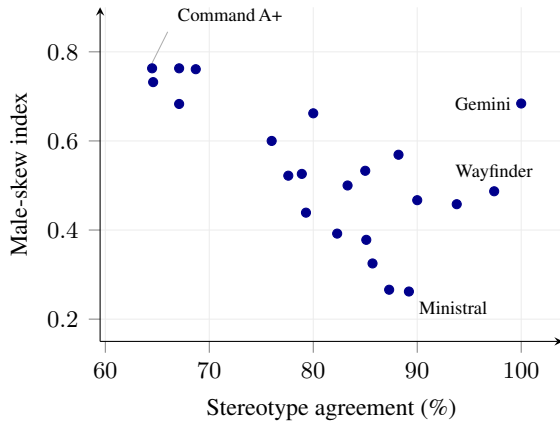


Figure 2: Two axes of bias on Set A (the 22 testable systems): stereotype agreement against the male-skew index. Every system lies above 50% agreement and above zero skew. The two are negatively correlated ($r = -0.60$): with the masculine side saturated, both are largely driven by how often a female-typed occupation is rendered as *she*, so a low agreement score (Command A+) does not indicate fairer behavior. Full per-system values are in Table 1.

Stereotype strength	items	all	male	female
60–70% one sex	8	34.4	100.0	21.2
70–80% one sex	21	71.0	99.5	35.6
80–90% one sex	20	79.3	100.0	65.2
90–100% one sex	33	94.1	100.0	77.3

Table 2: Set A stereotype agreement (%) by the strength of the occupational stereotype, pooled over the 22 testable systems and split by the direction of the label. Agreement on male-typed occupations is saturated in every band; the rise in the pooled figure comes entirely from the female-typed occupations.

leaves a mean of 20.0 of the 22 systems significant at a 10% error rate and 17.2 at 20%, with pooled agreement falling from 79.7% to 73.7% and then 67.9%. Directional concordance between the two labor markets would have to fall below roughly 70% before the pooled effect approached the no-bias line, which TÜİK (2024) and Gedikli (2020) make unlikely. Adversarially placed errors are harsher, and we state that bound in the Limitations.

The Set A effect is carried neither by a few items nor by one sentence frame. Leave-one-occupation-out moves pooled agreement only between 79.4% and 80.8%; a majority of systems follow the stereotype on 63 of the 82 occupations, and every committing system does so on 43. Pooled agreement across the ten template frames ranges from 64.7% to 90.2%, with every frame above the 50% base-

line. Because each occupation appears in exactly one frame, frame and occupation are confounded by design, so that spread is an upper bound on frame sensitivity rather than an estimate of it.

5.3 Contextual Disambiguation

The failure that dominated earlier evaluations, ignoring disambiguating context, has largely disappeared. Ten systems score a perfect 100% on Set B, including on all 54 stereotype-incongruent items, and 18 of 24 score at least 90%. A system that renders the ambiguous surgeon of Set A as *he* will still translate “Annem bir cerrah” (“My mother is a surgeon”) with *she* almost without exception. The stereotype therefore behaves as a prior that explicit evidence overrides, not as a filter that suppresses it. The expectation carried over from the WinoMT era (Stanovsky et al., 2019; Kocmi et al., 2020), in which reliance on stereotype over context was the headline result, no longer holds for current systems, which is itself a form of progress. What remains of the effect is confined to the weaker systems. GLM 4-9B falls from 85.7% on congruent items to 70.4% on incongruent ones ($\Delta S = +15.3$, Fisher $p = .18$), Mistral Medium 3.5 from 100% to 85.2% (+14.8, $p = .046$), and Tiny Aya Global from 92.9% to 81.5% (+11.4, $p = .21$). Only Mistral Medium reaches nominal significance, and with 28 congruent items the tests are underpowered, so we treat these as consistent trends rather than established effects.

When one of these weaker systems fails, the error is often more than a mispredicted pronoun; the system rewrites the gendered source word itself. Across the 46 kinship-cued Set B items, ten systems replaced the kinship term with one of the opposite gender in 31 outputs, and in 26 of those the replacement moved toward the occupation’s stereotype. None of the stronger systems did this; the behavior is limited to the weaker tail, led by Tiny Aya (seven times) and Mistral Medium 3.5 (five). A representative case is “Ablam bir mühendis” (“My older sister is an engineer”), which Tiny Aya renders as “My brother is an engineer. He is very well-liked by his colleagues.” Rather than resolving the pronoun from the kinship term, these systems overwrite the kinship term to fit the stereotype.

Where accuracy on Set B is uneven, it works against feminine referents. Most systems show little difference between masculine and feminine referents, but where a sizeable gap exists it runs the same way. Ministral 3-14B is 92.5% accu-

rate on masculine and 76.2% on feminine referents ($\Delta G = +16.3$, Fisher $p = .068$), and GLM 4-9B shows the same pattern (+8.6). The one clear reversal, Dubformer at -10.1 ($p = .10$), comes from its use of *they* on items whose gender is fixed, which our strict convention counts as an error, rather than from cross-gender substitution. No ΔG gap survives family-wise correction, so we note only that the direction is consistent, in line with Currey et al. (2022) and the weaker handling of feminine forms reported by Friðriksdóttir (2024).

5.4 Ambiguity Policy and Output Reliability

On what to do under ambiguity, the field divides into two groups. Neutralizers prefer singular *they*: VoxNexus_V1 uses *they* on 97.6% of Set A and 100% of Set C with no over-commitment, Dubformer neutralizes categorically, and Wayfinder (52.4%), GPT 5.5 (40.2%), and Lumen (37.8%) neutralize on a large share of Set A while keeping Set C over-commitment below 12%. Committers almost never abstain: Cohere CAT+, SalamandraTA, and Qingqiu-MT-9B produce no *they* at all and over-commit on every one of the neutral controls. The important point is that neutralization carries no measurable cost. Wayfinder, GPT 5.5, Lumen, and VoxNexus all reach 100% on Set B, the same as the best committers, because they switch to the correct gendered pronoun as soon as the context supplies one. On this suite singular *they* is a strictly better strategy, since it removes both stereotype bias and male skew on the ambiguous items without any loss of accuracy. Two qualifications are worth stating. Singular *they* is specific to English, and no equivalent is available when the target language marks gender grammatically (Rarrick et al., 2023; Friðriksdóttir, 2024). And neutralization can be taken too far: Dubformer applies it even where the context fixes gender, which costs it on Set B under our strict scoring (92.7%), a case we return to in the Ethical Considerations.

One frontier system fails as a translator rather than as a gender resolver. On 43.0% of items, Gemini 3.1 Pro returns output that cannot be scored as a single translation, either several alternative renderings labelled by register or a translation embedded in commentary about Turkish grammatical gender. Our normalization is deliberately charitable, since a single translation behind a chat preamble is unwrapped and scored, so 43.0% is a lower bound. We do not choose one of the alternatives on the system’s behalf, because that would hide the be-

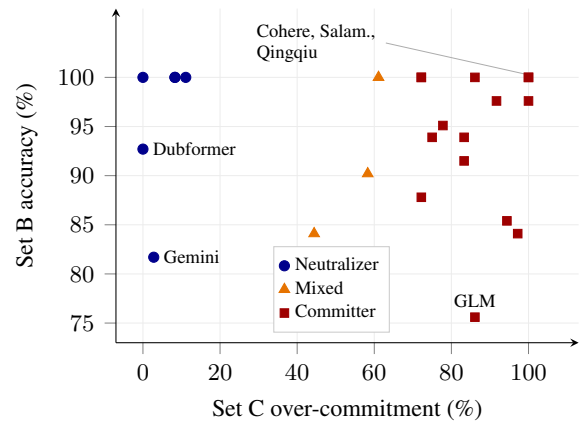


Figure 3: Ambiguity policy carries no accuracy cost. Systems that rarely over-commit on the neutral controls (left) reach the same near-ceiling Set B accuracy as heavy committers (right); Gemini’s lower accuracy reflects its malformed output, not gender errors.

havior. Where Gemini does commit it is accurate, with 100% stereotype agreement on its 19 committed Set A items and 2.8% over-commitment on Set C, so the problem lies in the reliability of the output format rather than in gender handling. No other system exceeds 0% malformed output, and the next largest reliability issue is the 6.0% hedge rate of GPT OSS 120B. This kind of failure has no counterpart in classical MT, and it is the reason our taxonomy treats malformed output as an outcome in its own right rather than as noise.

5.5 Interpretation

Any interpretation of these results has to respect the black-box setting. We see only the outputs of anonymized systems, with no access to weights, training data, or decoding settings, so what follows is a set of readings the behavior supports rather than established causes. The combination of universal stereotype tracking and universal male skew is consistent with systems that have learned the conditional probability of a pronoun given an occupation from training text which both mirrors real labor-market gender distributions and over-represents masculine pronouns overall; a decoding-side frequency effect would produce a similar pattern and cannot be excluded. The split between neutralizers and committers is consistent with a difference in training regime. Systems trained on aligned bitext have no abstention option, because human translators resolved the pronoun one way or the other, and the systems here that behave this way over-commit at 97% to 100%. Instruction-tuned models

plausibly learned to hedge under uncertainty and to express that hedge as singular *they*, and Gemini’s malformed output is consistent with chat-oriented tuning overriding the translation instruction when the input invites explanation. Because the systems are anonymized, we cannot confirm which belongs to which paradigm. The overall picture is nonetheless clear: for strong systems in 2026, gender bias in Turkish-to-English translation is no longer a problem of using context, since Set B performance is at or near ceiling, but a problem of ambiguity policy, of what a system should do when the gender of a referent cannot be recovered from the source.

6 Conclusion

We have presented `tr_gender_bias`, a 200-item reference-free Turkish-to-English test suite that turns the forced gender commitment imposed by the Turkish pronoun *o* into a controlled probe of gender bias. The suite pairs ambiguous items whose occupations carry a labor-statistics-grounded stereotype with context items that fix gender against that stereotype and with neutral controls, and it scores every output under a six-way abstention-aware taxonomy. Run on all 24 participating systems, it shows that stereotype-driven gender assignment under ambiguity is universal, significant for every testable system after correction, and consistently male-skewed, with no system ever skewing feminine. The two are closely linked rather than independent: commitments on male-typed occupations are masculine 830 times in 831, so what separates the systems is largely how often a female-typed occupation suffices to displace that default. At the same time, gender that the context states explicitly is now resolved almost perfectly, and the residual context failures appear only in the smaller models. Where masculine and feminine accuracy differ, the gap runs against feminine referents. The systems separate into committers and neutralizers, and neutralizing with singular *they* removes both stereotype bias and male skew on the ambiguous items at no cost in accuracy. Gender bias in current translation is therefore less a problem of resolving context than of deciding what to do when the source leaves gender open. The suite, its gold annotations, the deterministic generator, the evaluation script, and the robustness analysis of Section 5.2 are released under CC-BY-4.0.

Limitations

Our evaluation has several limits. Referent gender is approximated by the dominant gendered pronoun, which is sound only because each paragraph is built to contain a single human referent; it is not a general coreference method, and an output that introduces a second referent can be misattributed. The masculine, feminine, and neutral categories lean binary and do not probe neopronouns (cf. Friðriksdóttir, 2024). Referent-gender stereotypes are grounded in U.S. labor statistics, which approximate the association encoded in the largely English training data of these systems. Turkish labor data agree on the direction of that segregation (TÜİK, 2024; Gedikli, 2020) but are published only for broad occupation groups, so we cannot match the per-occupation magnitudes; the numerical strength of each stereotype is therefore a U.S. quantity, and the exact Turkish figure for a given occupation may differ. Section 5.2 bounds what this costs, but one bound there is worth stating plainly: errors placed adversarially, concentrated on exactly the items a system happened to get right, are far more damaging than random ones. Command A+, our lowest-agreement system, would lose nominal significance if only 2 of the 82 labels pointed the wrong way, although the median system tolerates 13. What is exposed at that margin is the universality of the effect across all 22 systems rather than the effect itself, whose pooled magnitude is far from the boundary. A further limit is one of scope. Our items deliberately strip away the discourse context, prior mentions, and world knowledge that in running text often resolve a referent long before *o* appears, so the rates we report describe behavior when ambiguity is isolated. They bound how often a system must guess rather than estimating how often it does guess in naturally occurring Turkish; the per-system ordering and the direction of each effect are the portable results. Because the systems are anonymized, every paradigm-level reading in our discussion is a behavioral hypothesis rather than an architectural claim. Finally, with only 28 congruent items, the Fisher tests for ΔS are underpowered.

Ethical Considerations

The suite measures how systems assign binary gendered pronouns under ambiguity. The “male” and “female” stereotype labels describe labor-market statistics; they are not claims about individuals or about who should hold an occupation. Rendering

an ambiguous surgeon as *she* is no more correct than rendering it as *he*, since both over-commit, and our ideal for an unknowable referent, singular *they*, asserts nothing about that referent, though we acknowledge that it is an English-specific normative choice (Rarrick et al., 2023; Friðriksdóttir, 2024). Our strict Set B convention, which counts *they* as an error when the context fixes gender, penalizes one defensible position, the categorical neutralization that Dubformer practices, so we report that behavior transparently rather than letting it disappear into the aggregate. All source texts are synthetic and contain no personal data. Finally, the per-system numbers measure a single phenomenon and should not be read as overall quality rankings.

Acknowledgments

We thank the WMT26 Test-Suite Subtask organizers for running the 24 participating systems on the suite and returning their outputs.

References

- Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. [Examining covert gender bias: A case study in Turkish and English machine translation models](#). In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Anna Currey, Maria Nadejde, Raghavendra 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.
- Steinunn Rut Friðriksdóttir. 2024. [The GenderQueer test suite](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 327–340, Miami, Florida, USA. Association for Computational Linguistics.
- Çiğdem Gedikli. 2020. [Occupational gender segregation in Turkey: The vertical and horizontal dimensions](#). *Journal of Family and Economic Issues*, 41(1):121–139.
- Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. [Gender coreference and bias evaluation at WMT 2020](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.
- Maja Popović and Ekaterina Lapshinova-Koltunski. 2025. [GENDER1PERSON: Test suite for estimating gender bias of first-person singular forms](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 800–822, Suzhou, China. Association for Computational Linguistics.
- Spencer Rarrick, Ranjita Naik, Sundar Poudel, and Vishal Chowdhary. 2023. [Evaluating gender bias in the translation of gender-neutral languages into English](#). *arXiv preprint arXiv:2311.08836*.
- 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.
- TÜİK. 2024. [İstatistiklerle kadın, 2024 \[women in statistics, 2024\]](#). Turkish Statistical Institute.