

Gender-Neutral Machine Translation Strategies in Practice

Hillary Dawkins

Isar Nejadgholi

Chi-kiu Lo 羅致翹

Digital Technologies Research Centre

National Research Council Canada (NRC-CNRC)

{hillary.dawkins, isar.nejadgholi, chikiu.lo}@nrc-cnrc.gc.ca

Abstract

Gender-inclusive machine translation (MT) should preserve gender ambiguity in the source to avoid misgendering and representational harms. While gender ambiguity often occurs naturally in notional gender languages such as English, maintaining that gender neutrality in grammatical gender languages is a challenge. Here we assess the sensitivity of 21 MT systems to the need for gender neutrality in response to gender ambiguity in three translation directions of varying difficulty. The specific gender-neutral strategies that are observed in practice are categorized and discussed. Additionally, we examine the effect of binary gender stereotypes on the use of gender-neutral translation. In general, we report a disappointing absence of gender-neutral translations in response to gender ambiguity. However, we observe a small handful of MT systems that switch to gender neutral translation using specific strategies, depending on the target language.

1 Introduction

Gender bias in language technologies has long been a concern for the community (see [Sun et al. \(2019\)](#) for an overview). However, machine translation (MT) stands out as a particularly important area of study due to its widespread use, its potential to foster global connections and uplift underserved communities, coupled with the unique technical challenges posed by grammatical gender differences across languages. Research on gender bias in machine translation often examines misgendering errors resulting from stereotypes or the default use of masculine forms ([Savoldi et al., 2021](#)). These errors contribute to both representational harms – such as the underrepresentation of female and non-binary individuals, or the limited depiction of women in certain roles ([Barclay and Sami, 2024](#);

[Stanovsky et al., 2019](#); [Troles and Schmid, 2021](#); [Sólmundsdóttir et al., 2022](#), *inter alia*) – and allocative harms, for example the increased burden on female users to manually correct machine-translated text ([Savoldi et al., 2024a](#)). Based on a human user study, [Dev et al. \(2021\)](#) found that such misgendering errors are particularly harmful for non-binary individuals. Here we are concerned with misgendering harms that arise, not due to stereotypes, but rather due to ambiguous or underspecified gender in the source text, a situation that occurs frequently in English.

One obvious solution is to use gender-inclusive language (i.e., language that encompasses all gender identities when the gender of a particular referent is unknown, or when gender is unimportant). However, the ease of this strategy varies by language. In notional gender languages, such as English, gender has minimal effects, primarily manifesting through gendered pronouns (e.g., he, himself) and relatively few gender-specific nouns (e.g., brother/sister, husband/wife, waiter/waitress, etc.). English also features some androcentric language (e.g., mankind) and, depending on the text domain and social norms, lends itself to the use of masculine generics when gender is irrelevant (e.g., “any driver in violation will lose his license”). It is easy to see then why gender inclusive language in English is relatively straightforward; the widely established use of “they/themselves” as a singular personal pronoun and the existence of in-vocabulary gender-neutral nouns (e.g., sister/brother → sibling, husband/wife → spouse, waiter/waitress → server, mankind → humankind) enable neutrality with minimal semantic or lexical perturbations. However, in grammatical gender languages (Spanish, Italian, French, Icelandic, Czech, etc.), multiple parts of speech typically need to agree with the gender of the referent (including nouns, verbs, and adjectives), and furthermore, grammatical gender may be restricted to binary (masculine/feminine)

© 2025 National Research Council Canada. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

cases. Therefore, gender neutrality can require more contrived rewrites, potentially using new or out-of-vocabulary words, and coherence may be sacrificed (Piergentili et al., 2023a). In the context of machine translation, a task *designed* to preserve meaning and coherence, a difficulty naturally arises in translation from lower to higher grammatical gender agreement contexts.

Gender neutral translation is the task of preserving gender neutrality in response to gender ambiguity in the source. Here, we study machine translation from English into grammatical gender languages, and assess MT systems’ *sensitivity* to the need for gender neutrality. We ask, if appropriate, low-barrier gender-neutral translation options exist in the target language, will modern MT systems use those options *in response* to gender ambiguity? We find that a deliberate form of gender ambiguity in English (using “they” as a singular personal pronoun) triggers a gender-neutral response in a small handful of surveyed MT systems, but overall, we observe a disappointing insensitivity to gender ambiguity. The results indicate much room for improvement, and we release the scoring code for the test set to enable future work at <https://github.com/hillary-dawkins/wmt24-gender-dialogue>.

2 The Task of Gender Neutral Translation

Based on the assertion that gender ambiguity should be preserved through translation to mitigate harm, **gender-neutral translation** (GNT) is an emergent challenge task to benchmark MT systems for this behaviour (Piergentili et al., 2023a). Through a careful analysis of existing gender-inclusive language guidelines for both English and Italian, and a survey of human participants, Piergentili et al. (2023a) define the desiderata of gender neutral translation: Neutrality should be maintained in the target if gender is ambiguous given the source, but conversely, gendered language should be used if it is known given the source. This makes gender neutrality a dynamic constraint, meaning MT systems should be *sensitive* to the need of it. In this study, we define a test set and metric to measure this sensitivity, termed the **gender-neutral response**, in a controlled way. Our test instances are templated to create pairs of inputs that differ only in whether the referent’s gender is known or unknown, and adjective translations in target gen-

dered languages are used to measure the sensitivity.

Prior studies in gender-neutral translation have been somewhat limited in scope; there has been a focus on translating gender-neutral pronouns (Cho et al., 2019; Barclay and Sami, 2024), including neo-pronouns (Lauscher et al., 2023), and gender-neutral nouns that are prone to masculine generic translation (Savoldi et al., 2023). These studies are focused on a first-order effect (i.e., the gender of the referent word itself – a pronoun or noun), whereas we are interested in second-order gender effects – the gender agreement of an adjective with its resolved referent. This is representative of the challenge that occurs when translating from English to grammatical gender languages.

Piergentili et al. (2023b) was the first, to our knowledge, to create a more robust, natural GNT benchmark (GeNTE), involving real in-the-wild inputs with reference translations (English → Italian). The test set contains both gender-ambiguous and gender-determined inputs paired with both gendered and gender-neutral reference translations. Using the GeNTE benchmark, Savoldi et al. (2024b) found that both neural MT systems and large language models (LLMs) struggle with gender-neutral translation, but we suggest this may be due to the difficulty of the test set: multiple gender agreements are needed – or should be avoided – in each output, and the target language, Italian, has only binary gender grammatical cases in its formal form. Furthermore, because the inputs are derived from real, naturally occurring language, it is hard to control for the existence (or difficulty) of a coherent gender-neutral translation.

Here, we take a step back and provide a somewhat easier and more controlled test set for GNT to compliment the GeNTE. Uniquely, the construction of the paired inputs, differing only in determined/ambiguous gender, allows us to measure a sensitivity to the need for gender neutrality – a prerequisite for GNT. In our setup, all gendered translations are measured via adjectives, and this provides several advantages. Firstly, the choice of gender declension after translation can be determined via dictionary lookup, rather than a learned classifier as in Piergentili et al. (2023b) and Savoldi et al. (2024b). The dictionary method allows for a more fine-grained analysis of specifically which gender-neutral strategies are being used in translation (e.g., a gender-neutral adjective vs. a noun phrase). Secondly, adjectives in our target languages typically have gender-neutral synonyms

available, and therefore the barrier to obtaining a gender-neutral translation is low (i.e., a human translator is not expected to struggle with this task). We include three target languages with different grammatical gender cases applied to adjectives:

1. Icelandic and Czech: Have a grammatical gender neuter case that does not usually apply to people.¹
2. Spanish: Has only masculine and feminine grammatical gender cases (in the standard/formal form of the language).

Although using templates to study gendered adjectives in isolation does not represent the breadth of real-world complexities, it does provide a starting point grounded in gender resolution that is typical of grammatical gendered languages.

3 Test Suite and Participating MT Systems

We make use of a test suite released as part of a shared challenge task at WMT24²: Gender Resolution in Speaker-Listener Dialogue Roles (Dawkins et al., 2024). All instances in the test suite involve spoken dialogue and meta-context surrounding the dialogue. Most instances involve two characters, and both characters may take on the speaker or listener role at times. Adjectives within spoken dialogue either refer to the speaker (self-referential) or the listener, and meta-context either resolves the gender of the adjectives’ referents, or leaves the gender ambiguous. As a minimal example: <“I’m/you’re stubborn”, I said to him.> either resolves the gender (you’re) or not (I’m). Examples of test suite instances supporting the current study are shown in Figure 1, and a full description of the test suite is provided in the Appendix.

The WMT24 shared task (Kocmi et al., 2024) garnered 40 unique MT systems operating in 11 translation directions, and the gender resolution test suite obtained translations from 21 unique MT systems in 3 language directions (EN → IS, CS, ES). Participating MT systems include commercial online systems, dedicated neural MT systems, and large language models (LLMs) prompted for MT. Throughout, results use the system names as provided by the shared task organizers (i.e., commercial online systems are anonymized). One key

challenge in gender-neutral translation, as highlighted by Piergentili et al. (2023a), is the need for paragraph-level inference to accurately determine gender. This consideration makes the WMT24 submissions particularly relevant and timely to study. Both the general MT shared task (formerly known as the News Translation task) and several challenge tasks placed an emphasis on paragraph-level translation. Among the 28 participating systems that specified their translation strategies, 17 reported using a paragraph-level or hybrid approach.

4 Observed Gender-Neutral Strategies

Over the entire test suite of 17,966 source adjectives, each translated into three target languages involving 21 unique MT systems, we observed 5 categories of gender-neutral translation strategies in practice:

1. **Gender-neutral adjectives** (N_1): The translated adjective takes the same inflection in either the feminine or masculine grammatical case.
2. **Neuter case adjectives** (N_2): For adjectives that do not have the same form for masculine and feminine cases, the grammatical gender neuter case is used if it exists (as in Icelandic and Czech).
3. **Other gender-neutral parts of speech** (N_3): Adjectives in the source are translated as a noun phrase or another part of speech without the need for gender agreement in the target language.
4. **English substitution** (N_4): The source adjective is directly copied in translation.
5. **New or alternative morphology** (N_5): New morphology is used to represent multiple possible variants of a gendered adjective in the target language. The observed variations mostly accommodate binary grammatical gender only³.

Table 1 provides an example of each gender-neutral translation strategy for each target language.

³Technically, alternative morphology with only binary options is not a fully gender-neutral strategy, but it is included here to measure an attempt at gender-neutral options in response to gender ambiguity. See Section 7 for further discussion.

¹Though dynamic social norms may permit this to varying degrees.

²<https://www2.statmt.org/wmt24/index.html>

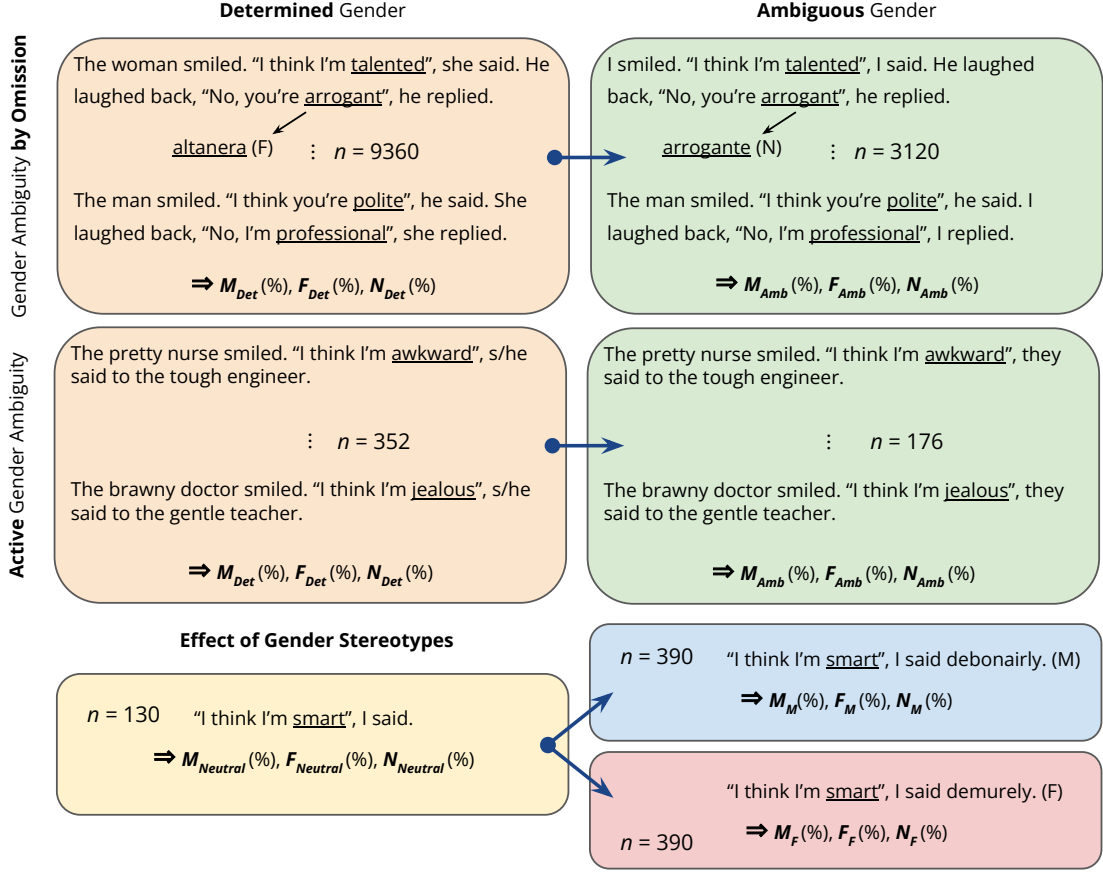


Figure 1: Examples of source inputs within the test suite. All inputs use adjectives to refer to either the speaker or the listener of known or unknown gender, given the context. Each determined gender template (orange) has a matching pair input with a small perturbation applied to make the gender ambiguous (green). The default masculine response $\Delta M = M_{Amb} - M_{Det}$ and the gender-neutral response $\Delta N = N_{Amb} - N_{Det}$ measure the effect of gender ambiguity on the resulting translation. Gender stereotyped adverbs (blue, red) are sometimes used to influence the assumed gender of the speaker (yellow). n represents the number of data points in each category.

5 Definitions and Metrics

We define the *gender neutrality* of an MT system on a specified subset of source adjective inputs as the proportion of gender-neutral translations observed in the output. *Gender-determined* inputs refer to source sentences that provide unambiguous gender information, such that the gender of the adjective referent is known. In contrast, *gender-ambiguous* inputs do not contain enough information to resolve the gender. For English source inputs, gender ambiguity naturally arises due to gender-neutral first- and second-person pronouns (I, you), and can also be induced through the use of “they” as a singular third-person pronoun. We refer to the former as *gender ambiguity by omission*, and the latter as *active gender ambiguity* since it involves a more deliberate choice to avoid “he/she” (and especially to avoid a default masculine form if gender is unknown or unimportant).

Here we are interested in how MT systems *respond* to a need to preserve gender ambiguity from source to target, and therefore we define metrics to measure the change in gender neutrality when source inputs switch from determined to ambiguous gender. Firstly, **baseline gender neutrality** (N_{Det}) measures the proportion of gender neutral translations in gender-determined cases, and provides a sense of how various MT systems make use of the gender-neutral strategies by default. Baseline gender neutrality is naturally expected to vary by target language due to variable availability, coherence, and ease of the gender-neutral strategies. When gender is made ambiguous, all else being equal, the **default masculine response** measures the increase in the proportion of masculine-form translations ($\Delta M = M_{Amb} - M_{Det}$), and the **gender-neutral response** measures the increase in the proportion of gender-neutral translations ($\Delta N = N_{Amb} - N_{Det}$). Refer to the examples

Source Adjective	Target Language	Gender-Neutral Strategies				
		N_1	N_2	N_3	N_4	N_5
fit	Spanish	fuerte	—	en forma	fit	musculos(o/a)
nonsensical	Czech	absurdní	nesmyslné	nemám smysl	nonsensical	nesmysln(ý/á)
cautious	Icelandic	varkár	varkárt	á varðbergi	cautious	huglítil(l)

Table 1: Examples of each gender-neutral translation type by target language.

in Figure 1 illustrating the move from determined to ambiguous gender in the source sentences.

Note that, depending on the context, either high baseline gender neutrality or high gender-neutral response (or some combination) could be defined as the ideal translation behaviour.⁴

6 Observations

6.1 Frequency of gender-neutral translations

Baseline gender neutrality is dominated by gender-neutral adjectives across languages and MT systems (refer to Appendix Table 6). As expected, the frequency varies by language due to a natural disparity in the availability of gender-neutral adjectives (on average, 39% of translations are gender neutral in Spanish, compared to 19% for both Czech and Icelandic)⁵. However, variance across MT systems is high within languages, suggesting that some systems strongly prefer the use of gender-neutral translations (e.g., in Icelandic, baseline gender neutrality ranges from 13% to 36%).

6.2 Strong masculine default response

When source sentences change from determined gender to ambiguous gender (Figure 1 orange to green), the **default masculine response** measures the increase in output masculine translations. The majority of surveyed MT systems have a strong tendency to default to a masculine translation under gender ambiguity in the source. This response occurs both when the gender is ambiguous by omission (refer to Table 4 ΔM), and when the gender is actively ambiguous (Table 5 ΔM).

Any response in the output translation when moving from determined to ambiguous gender implies a sensitivity to the need for gender agreement, al-

though, arguably, defaulting to a masculine form is not the best use of that awareness. The consequence of a default masculine response is fewer feminine translations under ambiguous gender, as shown by the change ΔF in Tables 4 and 5.

6.3 Gender-neutral sensitivity

We now arrive at the main result of the paper, which is to determine whether modern systems respond to the need for gender ambiguity in translation by employing the gender-neutral strategies. The so-called **gender-neutral response** measures the increase in the use of such strategies when source sentences are switched from determined to ambiguous gender.

Unfortunately, we observe *no significant gender-neutral response across any MT system or language* when the gender is made ambiguous by omission (I, you). All absolute changes in the proportion of gender-neutral translations are within $\pm 3\%$ (refer to Table 4 ΔN). This implies that although some systems prefer the use of gender-neutral translations, that preference is *not in response* to a need for gender neutrality.

However, we observe that more active gender ambiguity in the source (they) does trigger a gender-neutral response for a small handful of MT systems in Czech and Icelandic (refer to Table 2). Both within languages and across languages, the specific gender-neutral strategies driving the response differ. In Icelandic, the MT system with the strongest response (Claude-3.5) reacts to gender ambiguity by using alternative morphology ($\Delta N_5 = +22\%$), whereas Aya23 reacts by using the gender neuter case and other gender-neutral parts of speech. In Czech, all three systems with a significant response switch to using gender-neutral adjectives, a response that is never invoked in either Icelandic or Spanish. Furthermore, the response is inconsistent for specific systems across languages. For some gender-neutral strategies, this may be explained by divergent social norms encoded in the target language (e.g., Aya23 uses the neuter case in Icelandic, but not in Czech), but not so for other strategies.

⁴Although, based on the desiderata defined by Piergentili et al. (2023a), only high gender-neutral response would be ideal.

⁵Only 350 adjectives are included in the test suite – therefore these proportions are not likely representative of the availability of gender-neutral adjectives in the three target languages as a whole. This comparison only serves to illustrate that different baseline gender neutrality levels among target languages are not unexpected.

Lang.	System	$(M, F, N)_{Det}$	ΔN	ΔN_1	ΔN_2	ΔN_3	ΔN_4	ΔN_5
IS	Claude-3.5	(0.42, 0.36, 0.22)	0.180	-0.015	-0.012	-0.009	0.000	<u>0.215</u>
	Aya23	(0.50, 0.20, 0.30)	0.078	-0.003	<u>0.047</u>	0.039	-0.006	0.000
CS	CUNI-GA	(0.40, 0.40, 0.20)	0.081	<u>0.056</u>	0.000	0.017	0.009	0.000
	ONLINE-W	(0.39, 0.41, 0.19)	0.073	<u>0.070</u>	0.000	-0.005	0.007	0.000
	Unbabel-Tower	(0.40, 0.40, 0.19)	0.073	<u>0.050</u>	-0.003	0.022	0.003	0.000

Table 2: The 5 machine translation systems with a **non-zero gender-neutral response** observed in specific target languages – out of 21 systems in 3 language directions. The specific type of gender-neutral strategy contributing the most to the response is underlined. For example, when gender is known given the source, Claude-3.5 produces gender-neutral translations in Icelandic at a baseline rate of 22% (N_{Det}). When gender is made ambiguous in the source, this increases to 40% ($\Delta N = 0.180$), primarily through the use of alternative morphology (N_5).

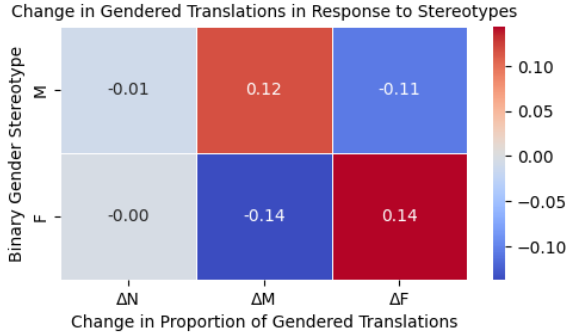


Figure 2: The average trade-off in the proportions of gendered translations (neutral, masculine, feminine) in response to the stereotyped adverbs (male or female), for the most affected MT systems (Table 3). Δ values represent absolute changes compared to the baseline levels with no stereotype present. Binary gender stereotypes affect the proportions of masculine and feminine translations, and have no effect on the neutral translations (Section 6.4).

For example, Claude-3.5 uses alternative morphology frequently in Icelandic (e.g., “(ur)” appending regular adjectives to denote both masculine and feminine cases), but seldom in the higher resource languages Czech or Spanish, despite an observed capacity to do so (e.g., “o/a” and “ý/á” are both observed infrequently). The full results for all systems and languages are shown in the Appendix Table 5.

6.4 Gender neutrality and stereotypes

Lastly, we observe that gender neutrality is not affected by binary gender stereotypes. That is, for each MT system and target language, the frequency of gender-neutral translations is consistent across source inputs (i) without any gender stereotypes (e.g., Figure 1 yellow box), and (ii) with binary gender stereotypes influencing the assumed gender of an adjective referent (e.g., Figure 1 blue and red

boxes).

Gender stereotypes are known to have a significant impact on the gender resolution of adjective referents, as measured by masculine vs. feminine adjective agreements. For example, “The nurse is talented” → “talentosa” (F) vs. “The doctor is talented” → “talentoso” (M). Though the trade-off between masculine and feminine cases in response to stereotypes is well known (see Dawkins et al. (2024) for a fuller analysis of stereotype effects within this test suite), here we include the proportion of neutral translations in the analysis. The results indicate that the trade-off occurs between the binary gender translations only, while the presence of the stereotype does not perturb gender-neutral translations (refer to the results in Table 3). Figure 2 illustrates the average trade-off between the gendered translations for the most affected systems. This implies that the base adjective choice is the first-order translation decision, and the gender agreement is second-order. This is consistent with the observed difficulty MT systems have in switching to a gender-neutral adjective in response to a need for gender neutrality; only three MT systems successfully use this strategy, and furthermore, it is only invoked under active gender ambiguity, as discussed in the previous section.

7 Discussion

Overall, the lack of alternative morphemes for unknown gender is disappointing; we observe alternative morphology (N_5) in only 4/21 systems (Dubformer, Claude-3.5, GPT4, and CommandR+), and usually at very low rates (less than 2% of gender-ambiguous inputs), with the exception of Claude-3.5 in the EN→IS translation direction under active gender ambiguity. While these LLM-based⁶

⁶Dubformer is a proprietary system; its architecture and training details are unknown.

System (Lang.)	$(M, F, N)_{Neutral}$	$(M, F, N)_{StereoM}$	$(M, F, N)_{StereoF}$	ΔG_{avg}	ΔN_{avg}
GPT-4 (IS)	(0.29, 0.48, 0.23)	(0.47, 0.32, 0.21)	(0.24, 0.54, 0.22)	0.119	-0.014
Claude-3.5 (CS)	(0.64, 0.18, 0.17)	(0.78, 0.06, 0.16)	(0.48, 0.35, 0.17)	0.147	-0.005
ONLINE-W (ES)	(0.48, 0.16, 0.36)	(0.51, 0.12, 0.36)	(0.28, 0.36, 0.36)	0.116	0.001

Table 3: Effect of binary **gender stereotypes** on the proportion of masculine (M), feminine (F), and gender-neutral translations (N). Subscripts denote the stereotype influencing the assumed gender of the adjective referent (either none, masculine, or feminine). The average effect on binary gender is denoted by ΔG_{avg} , and the average effect on the proportion of neutral translations is denoted by ΔN_{avg} . Here, the binary gender proportions are significantly impacted by stereotype cues (higher ΔG_{avg}), while the proportion of gender-neutral translations is roughly constant ($\Delta N_{avg} \approx 0$). The full results are shown in the Appendix Table 7.

H You will be asked to translate the following sentence from English into Icelandic. Translate: I am timid.

Ég er feiminn.

H You will be asked to translate the following sentence from English into Icelandic. Think step by step, what is the gender of the speaker. Translate: I am timid.

To translate "I am timid" to Icelandic, I need to consider the gender of the speaker since Icelandic adjectives change form based on gender.

Let me think about this step by step:

- In Icelandic, the adjective "timid" (feiminn) changes form based on gender:
 - For male speakers: "feiminn"
 - For female speakers: "feimin"
 - For neutral or unknown gender: "feimið"
- The pronoun "I" in Icelandic is "ég" for all genders.
- The verb "am" translates to "er" in Icelandic (same for all genders).

Since no gender information was specified in the sentence "I am timid," I should consider both possible translations:

For a male speaker: "Ég er feiminn"

For a female speaker: "Ég er feimin"

Without knowing the gender of the speaker, I'll provide both options:

If you're male: "Ég er feiminn"

If you're female: "Ég er feimin"

H The gender of the speaker is unknown. Avoid misgendering the speaker.

Since the gender of the speaker is unknown and I need to avoid misgendering, I should provide the translation that avoids assuming gender.

In Icelandic, when referring to oneself with an unknown gender, it's appropriate to provide both forms or explain the situation:

"Ég er feiminn/feimin"

This format acknowledges that Icelandic adjectives have gendered forms while avoiding assuming which is correct for the speaker. The speaker can then choose which form feels appropriate for them.

H The speaker might not identify as male or female.

You're absolutely right, and thank you for that important consideration. If the speaker doesn't identify as male or female, then I should provide an inclusive option.

In Icelandic, this presents a challenge since traditional grammar requires gendered adjective forms. For non-binary or gender-diverse speakers, some Icelandic speakers have started using the neuter form as a gender-neutral option:

"Ég er feimið"

This neuter form ("feimið") can be used as a gender-neutral alternative that doesn't assume a binary gender identity. While this usage is still evolving in Icelandic language practice, it represents an attempt to create more inclusive language options for non-binary individuals.

So a complete, inclusive translation would acknowledge:

"Ég er feiminn" (masculine)

"Ég er feimin" (feminine)

"Ég er feimið" (neuter/non-binary option)

(a) Claude-3.7

English → Spanish

I am timid × Soy tímida (feminine)

Soy tímido (masculine)

English → Icelandic

I am timid × Ég er hræddur

(b) Google Translate (Chrome)

You will be asked to translate a sentence from English into Spanish. Translate: I am muscular.

Soy musculoso.

You will be asked to translate a sentence from English into Spanish. Think step by step, what is the gender of the speaker? Translate: I am muscular.

The gender of the speaker is not explicitly stated in the sentence. In Spanish, adjectives must agree in gender with the subject.

- If the speaker is male: "Soy musculoso."
- If the speaker is female: "Soy musculosa."

Since we don't have gender information, both translations are possible.

The speaker may not be either female or male.

Good point! If the speaker does not identify as male or female, a gender-neutral option would be preferred. While Spanish traditionally uses gendered adjectives, some gender-neutral alternatives exist:

- "Soy musculoso." (Using the -o ending, which some non-binary Spanish speakers prefer.)
- "Tengo músculos." (Avoids gendered adjectives while conveying the same idea.)

Would you like a specific approach based on the context?

What are some synonyms for "muscular" that would be gender-neutral adjectives in Spanish?

Great question! In Spanish, most adjectives are gendered (-o for masculine, -a for feminine), but some remain neutral. Here are some gender-neutral synonyms for "muscular":

- "Fuerte" (strong)
- "Robusto" (robust, using the non-binary -e form)
- "Definido" (defined, using the -e form)
- "Atlético" (athletic, with the -e form)
- "Fornido" (stocky, strong-built, with the -e form)

Would you like me to adjust the translation with one of these?

Provide the best gender-neutral translation.

The best gender-neutral translation would be:

"Soy fuerte."

"Fuerte" is already a naturally gender-neutral adjective in Spanish, making it widely understood and grammatically correct. It conveys the idea of being muscular without needing gendered endings.

(c) ChatGPT

Figure 3: Examples of common online translation systems in response to gender ambiguity.

translation systems have the capacity and freedom to use gender-neutral morphology, we find that this is not usually invoked in practice, even for simple adjectives. Furthermore, the new morphemes that we do observe typically conform to a binary gender assumption (e.g., “o/a” for masculine and feminine cases), with the sole exception of Dubformer, which uses “@” in place of “o/a” in Spanish.

Investigating further, we observe that popular solutions for casual MT users have a tendency to push binary gender solutions – if gender ambiguity is acknowledged at all. Referring to the examples shown in Figure 3, Google Translate tends to offer default masculine translations for lower resource languages (Icelandic), and masculine/feminine variants for higher resource languages (Spanish), but never a gender-neutral variant of the gendered adjectives. Both Claude-3.7 and ChatGPT provide masculine default translations given a simple translate instruction. However, chain-of-thought prompting can successfully be used to force the LLMs to consider the gender of the adjective’s referent. Binary gender alternatives are produced when LLMs are instructed to consider gender, and gender-neutral strategies are used only when explicitly reminded that gender is not binary. When gender neutrality is requested, the specific strategies can vary. The examples show Claude-3.7 using the neuter case, and ChatGPT using an alternative part of speech (e.g., switching “muscular” for “having muscles”). It does appear that prompting can be used to steer the LLMs towards a preferred gender-neutral strategy (e.g., asking ChatGPT to suggest gender-neutral synonyms for “muscular” produces “fuerte (strong)”).

Large language models in combination with prompting strategies may be a promising direction for gender-neutral machine translation, but some limitations remain. Although LLMs have the power to provide these nuanced explanations with multiple translation options, they often fail to give *usable* translation options given the context. That is, providing an explanation with feminine and masculine options may be a partially *correct* response, but it is neither *complete* (lacking non-binary options), nor *useful* given the information in the source (lacking gender neutrality given gender ambiguity). Additionally, the explanation provided by the LLM must be distilled to a singular output translation if used at scale. Ultimately, LLM-based methods should be assessed based on a single output translation based on the (limited) information

provided by the source input.

Given that machine translation systems can be steered toward a particular strategy, it becomes important to consider what the ideal gender-neutral translation should be. The availability of gender neutral adjectives, or other parts of speech, naturally varies by target language. Forcing gender neutrality can decrease translation quality, both in terms of coherence in the target language and faithfulness to the source input. Should gender neutrality in translation only be enforced in response to gender ambiguity (high gender neutral response), or should it be enforced for all inputs (high baseline gender neutrality)? A high gender neutral response indicates less overall intervention, but may introduce a disparity in translation quality for non-binary genders.

Finally, it should be noted that the proposed metrics capture a certain aspect of gender-inclusive machine translation, but do not cover all aspects. For example, gender stereotypes can negatively affect translations while maintaining gender neutrality, and this effect is not captured by either the baseline gender neutrality or by the gender neutral response. We observe mismatched adjective translations depending on binary gender stereotypes (e.g. “thick” has divergent translation as either “dumb” (M) or “fat” (F), likewise “modest” → “humble”(M)/“covered”(F), etc.). If these divergent translations are each gender-neutral in the target language, the stereotype effect will be obscured. Therefore, gender neutrality should be considered as just one branch of ideal gender-inclusive machine translation.

8 Conclusion

Gender-neutral translation is a path toward gender-inclusive machine translation that helps prevent misgendering and representational biases. However, languages with grammatical gender pose significant technical challenges. A prerequisite for MT systems is the ability to recognize when gender neutrality should be applied. In a large-scale analysis of 21 MT systems, we evaluated their sensitivity to gender ambiguity in the source text. While a masculine default response is often observed, the gender-neutral response is lacking. That is, MT systems struggle to switch to gender-neutral strategies, despite the availability of such strategies in the target languages (e.g., grammatical gender neuter cases, abundant gender-neutral in-vocabulary al-

ternatives, and an observed capacity to use alternative morphology). Our findings highlight the need for further research in this challenging subtask of gender-inclusive translation. Perhaps encouragingly, a small handful of systems did demonstrate a non-zero gender-neutral response, though triggered only by active gender ambiguity. Future work may examine the properties of these systems that enabled such a response, in order to strengthen the effect in other contexts. Additionally, future work might examine the utility of the templated inputs used here, differing only in gender ambiguity, as a training set to improve gender-neutral translation in more complex scenarios, such as those found in the GeNTE benchmark.

References

- Peter J Barclay and Ashkan Sami. 2024. [Investigating markers and drivers of gender bias in machine translations](#). In *2024 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 455–464.
- 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.
- Hillary Dawkins, Isar Nejadgholi, and Chi-Kiu Lo. 2024. [WMT24 test suite: Gender resolution in speaker-listener dialogue roles](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 307–326, Miami, Florida, USA. Association for Computational Linguistics.
- Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. [Harms of gender exclusivity and challenges in non-binary representation in language technologies](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinhórfur Steingrímsson, and Vilém Zouhar. 2024. [Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. 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.
- Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023a. [Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges](#). In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland. European Association for Machine Translation.
- Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023b. [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.
- 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, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. [Test suites task: Evaluation of gender fairness in MT with MuST-SHE and INES](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore. Association for Computational Linguistics.
- Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024a. [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.
- Beatrice Savoldi, Andrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024b. [A prompt response to the demand for automatic gender-neutral translation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 256–267, St. Julian’s, Malta. Association for Computational Linguistics.
- Agnes Sólmundsdóttir, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir, and Anton Ingason. 2022. [Mean machine translations: On gender bias in Icelandic machine translations](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3113–3121, Marseille, France. European Language Resources Association.
- 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.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. [Mitigating gender bias in natural language processing: Literature review](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Jonas-Dario Troles and Ute Schmid. 2021. [Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online. Association for Computational Linguistics.

A Test Suite Details and Full Results

The full “gender resolution in literary dialogue settings” test suite contains various types of source inputs, each containing dialogue and meta-context surrounding the dialogue. The meta-context may include gender-stereotyped character descriptions and adverbs that control the manner of speaking. In all cases, adjectives within spoken dialogue refer to characters in the meta-context. The challenge of the test suite is to use the meta-context to correctly resolve each adjective’s referent, as measured through gender agreement.

For our purposes, we take subsets of the test suite that can be paired to create (determined, ambiguous)-gender inputs, while all other details remain constant. Test suite inputs that support the study of gender ambiguity by omission take the form:

The {woman, man} smiled. “I think {I’m, you’re} A_1 ,” {she, he} said.

{He, She} laughed back. “No, [{you’re, I’m} not A_1 , but] {you are, I am} A_2 ,” {he, she} replied.

(1)

The {man, woman} smiled. “I think I’m A_1 and you’re A_2 ,” {he, she} said.

{He, She} laughed back. “No, you’re A_3 , but I’m A_4 ,” {he, she} replied.

(2)

{I, The wo/man} smiled. “I think {I’m, you’re} A_1 ,” {I, s/he} said.

{S/he, I} laughed back. “No, [{you’re, I’m} not A_1 , but] {you are, I am} A_2 ,” {s/he, I} replied.

(3)

{I, The wo/man} smiled. “I think I’m A_1 and you’re A_2 ,” {I, s/he} said.

{S/He, I} laughed back. “No, you’re A_3 , but I’m A_4 ,” {s/he, I} replied.

(4)

where adjectives are denoted by A_i , curly braces denote template variables, and square brackets indicate optional text in templates (1) and (3). Adjectives in templates (1) and (2) are always gender-determined ($n_1^{Det} = 2400$, $n_2^{Det} = 3840$). A perturbation in template (1) makes the all adjectives gender-ambiguous half of the time in template (3) ($n_3^{Det} = 1200$, $n_3^{Amb} = 1200$), and similarly, a perturbation in template (2) makes half of the adjectives gender-ambiguous in every instance of template (4) ($n_4^{Det} = 1920$, $n_4^{Amb} = 1920$). In all cases, binary gender and the position of the first-person speaker, as applicable to the template, are balanced across the test suite. The full results for the proportions of gendered adjective translations in the case of determined vs. ambiguous gender, using these template pairs, are shown in Table 4. The results are macro-averaged over each template type as applicable. The baseline levels of gender-neutral translations, broken down by type (Section 4), in the gender-determined cases are shown in Table 6.

Test suite instances that support the study of active gender ambiguity take the form:

The C_g smiled. “I think I’m A ,”
{he, she, they} said to the $C_{\bar{g}}$.

(5)

where $(C_g, C_{\bar{g}})$ pairs denote binary gender stereotyped character descriptions,

$$C_g = a_g occ_g, \quad (6)$$

where a_g is a gender-stereotyped adjective, and occ_g is a matching gender-stereotyped occupation (e.g. “pretty nurse” or “strong doctor”). Template instances with “he” or “she” create gender-determined adjectives ($n_5^{Det} = 352$), while instances with “they” create gender-ambiguous adjectives ($n_5^{Amb} = 176$). The binary gender stereotypes

are balanced across {he, she, they} instances. The full results using these template pairs are shown in Table 5.

Finally, the effect of binary gender stereotypes on the gender neutrality is studied using test suite instances of the form:

$$\text{“I think I’m } A,\text{” I said [} \textit{adverb} \text{]}. \quad (7)$$

where *adverb* is optionally used to control the manner of speaking and aligns with a socially held stereotype about binary gender (e.g., gently vs. brusquely). In total, $n_7^{\textit{neutral}} = 130$ instances without any adverb are paired with $n_7^M = n_7^F = 390$ stereotyped instances, balanced by binary gender. Dawkins et al. (2024) find that stereotyped manner of speaking significantly impacts the choice of gender declensions between feminine and masculine forms in translation, and here we include the levels of gender-neutral translations to observe if neutrality is affected. The full results are shown in Table 7.

Lang.	System	$(M, F, N)_{Det}$	$(M, F, N)_{Amb}$	ΔM	ΔF	ΔN
IS	AMI	(0.56, 0.31, 0.13)	(0.67, 0.19, 0.14)	0.110	-0.113	0.003
	Aya23	(0.39, 0.25, 0.36)	(0.46, 0.21, 0.33)	0.066	-0.037	-0.029
	Claude-3.5	(0.44, 0.42, 0.15)	(0.50, 0.33, 0.17)	0.068	-0.085	0.017
	Dubformer	(0.62, 0.23, 0.16)	(0.69, 0.13, 0.18)	0.073	-0.099	0.026
	GPT-4	(0.41, 0.41, 0.18)	(0.34, 0.49, 0.17)	-0.068	0.080	-0.012
	IKUN	(0.33, 0.48, 0.19)	(0.24, 0.56, 0.20)	-0.084	0.077	0.007
	IOL-Research	(0.46, 0.36, 0.18)	(0.65, 0.19, 0.16)	0.195	-0.180	-0.015
	Llama3-70B	(0.47, 0.31, 0.21)	(0.56, 0.24, 0.20)	0.086	-0.074	-0.012
	ONLINE-A	(0.62, 0.20, 0.19)	(0.72, 0.10, 0.18)	0.100	-0.092	-0.008
	ONLINE-B	(0.55, 0.28, 0.16)	(0.65, 0.18, 0.17)	0.103	-0.108	0.005
	ONLINE-G	(0.57, 0.15, 0.28)	(0.61, 0.12, 0.27)	0.043	-0.035	-0.008
	TranssionMT	(0.55, 0.29, 0.16)	(0.66, 0.18, 0.17)	0.104	-0.110	0.006
	Unbabel-Tower70B	(0.46, 0.39, 0.15)	(0.57, 0.29, 0.14)	0.106	-0.097	-0.009
CS	Aya23	(0.42, 0.39, 0.20)	(0.31, 0.51, 0.18)	-0.109	0.125	-0.016
	CUNI-DocTransformer	(0.44, 0.38, 0.18)	(0.66, 0.17, 0.17)	0.211	-0.205	-0.006
	CUNI-GA	(0.45, 0.35, 0.20)	(0.53, 0.29, 0.18)	0.076	-0.063	-0.013
	CUNI-MH	(0.37, 0.43, 0.21)	(0.22, 0.60, 0.18)	-0.146	0.171	-0.025
	Claude-3.5	(0.42, 0.42, 0.16)	(0.66, 0.19, 0.15)	0.234	-0.222	-0.012
	CommandR+	(0.40, 0.42, 0.18)	(0.47, 0.37, 0.16)	0.067	-0.048	-0.019
	GPT-4	(0.45, 0.42, 0.13)	(0.61, 0.25, 0.13)	0.163	-0.163	-0.000
	IKUN	(0.43, 0.32, 0.25)	(0.58, 0.18, 0.24)	0.149	-0.139	-0.010
	IOL-Research	(0.43, 0.42, 0.16)	(0.74, 0.11, 0.15)	0.315	-0.309	-0.006
	Llama3-70B	(0.45, 0.35, 0.19)	(0.77, 0.03, 0.19)	0.322	-0.320	-0.002
	NVIDIA-NeMo	(0.55, 0.28, 0.18)	(0.57, 0.27, 0.16)	0.023	-0.010	-0.013
	ONLINE-A	(0.44, 0.39, 0.18)	(0.36, 0.49, 0.16)	-0.079	0.100	-0.021
	ONLINE-B	(0.58, 0.25, 0.18)	(0.75, 0.07, 0.18)	0.175	-0.180	0.005
	ONLINE-G	(0.54, 0.26, 0.19)	(0.71, 0.12, 0.18)	0.163	-0.146	-0.017
	ONLINE-W	(0.44, 0.40, 0.17)	(0.37, 0.47, 0.16)	-0.065	0.075	-0.010
	SCIR-MT	(0.45, 0.38, 0.18)	(0.47, 0.36, 0.17)	0.023	-0.014	-0.009
	TranssionMT	(0.46, 0.36, 0.18)	(0.44, 0.39, 0.16)	-0.022	0.033	-0.011
	Unbabel-Tower70B	(0.44, 0.37, 0.20)	(0.49, 0.33, 0.18)	0.056	-0.042	-0.014
ES	Aya23	(0.34, 0.26, 0.40)	(0.42, 0.22, 0.37)	0.077	-0.047	-0.030
	Claude-3.5	(0.31, 0.30, 0.38)	(0.45, 0.18, 0.37)	0.141	-0.127	-0.014
	CommandR+	(0.33, 0.31, 0.36)	(0.26, 0.39, 0.36)	-0.072	0.075	-0.003
	Dubformer	(0.41, 0.23, 0.37)	(0.55, 0.09, 0.36)	0.141	-0.135	-0.006
	GPT-4	(0.33, 0.29, 0.38)	(0.56, 0.06, 0.37)	0.235	-0.230	-0.005
	IKUN	(0.29, 0.31, 0.40)	(0.28, 0.32, 0.40)	-0.013	0.018	-0.005
	IOL-Research	(0.34, 0.28, 0.38)	(0.45, 0.17, 0.38)	0.111	-0.115	0.004
	Llama3-70B	(0.35, 0.27, 0.38)	(0.56, 0.09, 0.36)	0.208	-0.186	-0.022
	MSLC	(0.38, 0.15, 0.47)	(0.42, 0.12, 0.46)	0.044	-0.036	-0.008
	ONLINE-A	(0.41, 0.21, 0.38)	(0.51, 0.11, 0.38)	0.105	-0.101	-0.004
	ONLINE-B	(0.38, 0.22, 0.40)	(0.50, 0.11, 0.39)	0.118	-0.104	-0.014
	ONLINE-G	(0.40, 0.20, 0.39)	(0.47, 0.14, 0.38)	0.069	-0.060	-0.009
	ONLINE-W	(0.33, 0.30, 0.37)	(0.40, 0.23, 0.37)	0.066	-0.069	0.003
	TranssionMT	(0.39, 0.21, 0.40)	(0.51, 0.11, 0.38)	0.120	-0.101	-0.019
	Unbabel-Tower70B	(0.32, 0.29, 0.39)	(0.37, 0.23, 0.40)	0.046	-0.052	0.006

Table 4: **Responses to Gender Ambiguity by Omission:** A comparison of the proportion of masculine (M), feminine (F), and gender-neutral translations (N) when the gender of the adjective referent is known (subscript Det) vs. when it is ambiguous by omission in the source via “I” or “you” (subscript Amb). The default masculine response is denoted by ΔM , and the gender-neutral response is denoted by ΔN .

Lang.	System	$(M, F, N)_{Det}$	$(M, F, N)_{Amb}$	ΔM	ΔF	ΔN
IS	AMI	(0.46, 0.43, 0.11)	(0.72, 0.15, 0.13)	0.262	-0.282	0.020
	Aya23	(0.50, 0.20, 0.30)	(0.41, 0.21, 0.38)	-0.092	0.014	0.078
	Claude-3.5	(0.42, 0.36, 0.22)	(0.51, 0.09, 0.40)	0.087	-0.267	0.180
	Dubformer	(0.68, 0.14, 0.19)	(0.72, 0.11, 0.17)	0.042	-0.026	-0.016
	GPT-4	(0.41, 0.42, 0.18)	(0.51, 0.28, 0.22)	0.100	-0.141	0.041
	IKUN	(0.37, 0.44, 0.19)	(0.41, 0.37, 0.22)	0.038	-0.07	0.032
	IOL-Research	(0.43, 0.38, 0.19)	(0.68, 0.11, 0.21)	0.250	-0.273	0.023
	Llama3-70B	(0.46, 0.33, 0.22)	(0.55, 0.21, 0.24)	0.092	-0.118	0.026
	ONLINE-A	(0.46, 0.35, 0.20)	(0.73, 0.09, 0.19)	0.270	-0.258	-0.012
	ONLINE-B	(0.45, 0.40, 0.15)	(0.51, 0.33, 0.16)	0.064	-0.073	0.009
	ONLINE-G	(0.43, 0.33, 0.23)	(0.61, 0.16, 0.23)	0.177	-0.177	0.000
	TranssionMT	(0.45, 0.40, 0.15)	(0.50, 0.34, 0.16)	0.052	-0.061	0.009
	Unbabel-Tower70B	(0.44, 0.40, 0.16)	(0.64, 0.19, 0.16)	0.199	-0.206	0.007
CS	Aya23	(0.49, 0.33, 0.18)	(0.60, 0.23, 0.17)	0.114	-0.103	-0.011
	CUNI-DocTransformer	(0.40, 0.40, 0.19)	(0.53, 0.22, 0.25)	0.128	-0.187	0.059
	CUNI-GA	(0.40, 0.40, 0.20)	(0.53, 0.19, 0.28)	0.128	-0.209	0.081
	CUNI-MH	(0.36, 0.38, 0.26)	(0.43, 0.29, 0.29)	0.071	-0.094	0.023
	Claude-3.5	(0.41, 0.43, 0.16)	(0.66, 0.13, 0.22)	0.246	-0.303	0.057
	CommandR+	(0.39, 0.40, 0.21)	(0.69, 0.10, 0.22)	0.297	-0.3	0.003
	GPT-4	(0.45, 0.41, 0.14)	(0.69, 0.14, 0.17)	0.246	-0.268	0.022
	IKUN	(0.44, 0.35, 0.21)	(0.55, 0.22, 0.23)	0.107	-0.125	0.018
	IOL-Research	(0.40, 0.40, 0.20)	(0.61, 0.20, 0.19)	0.209	-0.201	-0.008
	Llama3-70B	(0.44, 0.36, 0.20)	(0.65, 0.13, 0.22)	0.215	-0.233	0.018
	NVIDIA-NeMo	(0.40, 0.40, 0.19)	(0.62, 0.16, 0.22)	0.216	-0.243	0.027
	ONLINE-A	(0.41, 0.41, 0.18)	(0.52, 0.30, 0.19)	0.108	-0.111	0.003
	ONLINE-B	(0.41, 0.41, 0.19)	(0.76, 0.02, 0.22)	0.349	-0.384	0.035
	ONLINE-G	(0.39, 0.39, 0.22)	(0.72, 0.04, 0.24)	0.328	-0.351	0.023
	ONLINE-W	(0.39, 0.41, 0.19)	(0.48, 0.25, 0.27)	0.085	-0.158	0.073
	SCIR-MT	(0.40, 0.40, 0.20)	(0.56, 0.24, 0.20)	0.159	-0.159	0.000
	TranssionMT	(0.40, 0.40, 0.20)	(0.57, 0.20, 0.22)	0.170	-0.198	0.028
	Unbabel-Tower70B	(0.40, 0.40, 0.19)	(0.48, 0.25, 0.27)	0.077	-0.15	0.073
ES	Aya23	(0.32, 0.26, 0.42)	(0.36, 0.22, 0.43)	0.039	-0.044	0.005
	Claude-3.5	(0.41, 0.22, 0.37)	(0.51, 0.11, 0.38)	0.098	-0.105	0.007
	CommandR+	(0.34, 0.29, 0.36)	(0.56, 0.09, 0.35)	0.217	-0.205	-0.012
	Dubformer	(0.39, 0.24, 0.38)	(0.52, 0.11, 0.37)	0.137	-0.126	-0.011
	GPT-4	(0.44, 0.18, 0.38)	(0.56, 0.05, 0.39)	0.120	-0.13	0.010
	IKUN	(0.36, 0.21, 0.43)	(0.46, 0.15, 0.39)	0.095	-0.06	-0.035
	IOL-Research	(0.34, 0.29, 0.38)	(0.49, 0.11, 0.40)	0.153	-0.18	0.027
	Llama3-70B	(0.40, 0.21, 0.39)	(0.51, 0.13, 0.35)	0.112	-0.08	-0.032
	MSLC	(0.37, 0.22, 0.41)	(0.53, 0.10, 0.37)	0.157	-0.119	-0.038
	ONLINE-A	(0.32, 0.27, 0.40)	(0.56, 0.08, 0.36)	0.235	-0.197	-0.038
	ONLINE-B	(0.30, 0.27, 0.43)	(0.54, 0.06, 0.39)	0.241	-0.206	-0.035
	ONLINE-G	(0.31, 0.27, 0.42)	(0.51, 0.09, 0.39)	0.201	-0.179	-0.022
	ONLINE-W	(0.39, 0.24, 0.38)	(0.51, 0.12, 0.37)	0.121	-0.115	-0.006
	TranssionMT	(0.31, 0.27, 0.41)	(0.55, 0.08, 0.37)	0.232	-0.191	-0.041
	Unbabel-Tower70B	(0.30, 0.31, 0.38)	(0.43, 0.17, 0.41)	0.120	-0.143	0.023

Table 5: **Responses to Active Gender Ambiguity:** A comparison of the proportion of masculine (M), feminine (F), and gender-neutral translations (N) when the gender of the adjective referent is known (subscript Det) vs. when it is actively ambiguous via “they” (subscript Amb). The default masculine response is denoted by ΔM , and the gender-neutral response is denoted by ΔN .

Lang.	System	N_{Det}	N_1	N_2	N_3	N_4	N_5
IS	AMI	0.133	0.125	0.003	0.003	0.002	0.000
	Aya23	0.360	0.186	0.122	0.030	0.022	0.000
	Claude-3.5	0.149	0.132	0.009	0.004	0.000	0.003
	Dubformer	0.158	0.111	0.037	0.002	0.003	0.005
	GPT-4	0.177	0.144	0.030	0.002	0.001	0.000
	IKUN	0.191	0.178	0.008	0.002	0.002	0.000
	IOL-Research	0.177	0.147	0.012	0.001	0.016	0.000
	Llama3-70B	0.213	0.147	0.046	0.011	0.009	0.000
	ONLINE-A	0.185	0.141	0.008	0.004	0.033	0.000
	ONLINE-B	0.163	0.155	0.004	0.004	0.000	0.000
	ONLINE-G	0.279	0.153	0.039	0.001	0.086	0.000
	TranssionMT	0.163	0.154	0.004	0.005	0.000	0.000
	Unbabel-Tower70B	0.151	0.139	0.007	0.004	0.001	0.000
CS	Aya23	0.196	0.132	0.001	0.044	0.019	0.000
	CUNI-DocTransformer	0.180	0.089	0.001	0.065	0.025	0.000
	CUNI-GA	0.197	0.088	0.001	0.089	0.019	0.000
	CUNI-MH	0.206	0.099	0.000	0.087	0.019	0.000
	Claude-3.5	0.161	0.095	0.000	0.048	0.018	0.000
	CommandR+	0.180	0.090	0.000	0.068	0.022	0.000
	GPT-4	0.132	0.084	0.000	0.031	0.018	0.000
	IKUN	0.246	0.129	0.003	0.089	0.026	0.000
	IOL-Research	0.155	0.082	0.000	0.057	0.017	0.000
	Llama3-70B	0.194	0.145	0.003	0.020	0.026	0.000
	NVIDIA-NeMo	0.175	0.108	0.002	0.047	0.019	0.000
	ONLINE-A	0.177	0.094	0.002	0.060	0.021	0.000
	ONLINE-B	0.179	0.098	0.001	0.060	0.021	0.000
	ONLINE-G	0.193	0.127	0.000	0.039	0.027	0.000
	ONLINE-W	0.169	0.088	0.000	0.063	0.018	0.000
	SCIR-MT	0.177	0.109	0.000	0.046	0.022	0.000
	TranssionMT	0.175	0.100	0.002	0.054	0.019	0.000
	Unbabel-Tower70B	0.196	0.106	0.000	0.073	0.017	0.000
ES	Aya23	0.396	0.367	0.000	0.021	0.009	0.000
	Claude-3.5	0.383	0.372	0.000	0.011	0.000	0.000
	CommandR+	0.359	0.341	0.000	0.016	0.002	0.000
	Dubformer	0.367	0.345	0.000	0.012	0.005	0.005
	GPT-4	0.379	0.362	0.000	0.015	0.002	0.000
	IKUN	0.401	0.375	0.000	0.021	0.004	0.000
	IOL-Research	0.379	0.360	0.000	0.018	0.001	0.000
	Llama3-70B	0.378	0.353	0.000	0.020	0.005	0.000
	MSLC	0.471	0.392	0.000	0.044	0.035	0.000
	ONLINE-A	0.382	0.340	0.000	0.036	0.005	0.000
	ONLINE-B	0.403	0.351	0.000	0.047	0.006	0.000
	ONLINE-G	0.393	0.345	0.000	0.043	0.006	0.000
	ONLINE-W	0.367	0.321	0.000	0.039	0.007	0.000
	TranssionMT	0.399	0.350	0.000	0.043	0.006	0.000
	Unbabel-Tower70B	0.389	0.360	0.000	0.029	0.001	0.000

Table 6: The **baseline gender neutrality** (N_{Det}) by type ($N_{\{1...5\}}$) for all translation systems for the three target languages.

	System	$(M, F, N)_{Neutral}$	$(M, F, N)_{Stereom}$	$(M, F, N)_{StereoF}$	ΔG_{avg}	ΔN_{avg}
IS	AMI	(0.69, 0.15, 0.16)	(0.75, 0.09, 0.16)	(0.74, 0.09, 0.17)	0.001	0.005
	Aya23	(0.44, 0.26, 0.30)	(0.45, 0.20, 0.34)	(0.38, 0.27, 0.35)	0.010	0.048
	Claude-3.5	(0.56, 0.12, 0.32)	(0.70, 0.05, 0.25)	(0.52, 0.17, 0.31)	0.093	-0.037
	Dubformer	(0.70, 0.10, 0.20)	(0.65, 0.14, 0.20)	(0.62, 0.15, 0.23)	0.003	0.017
	GPT-4	(0.29, 0.48, 0.23)	(0.47, 0.32, 0.21)	(0.24, 0.54, 0.22)	0.119	-0.014
	IKUN	(0.15, 0.56, 0.28)	(0.24, 0.48, 0.28)	(0.15, 0.58, 0.27)	0.051	-0.010
	IOL-Research	(0.56, 0.17, 0.27)	(0.60, 0.12, 0.27)	(0.53, 0.19, 0.27)	0.033	0.005
	Llama3-70B	(0.57, 0.17, 0.27)	(0.65, 0.11, 0.24)	(0.50, 0.23, 0.28)	0.070	-0.008
	ONLINE-A	(0.63, 0.09, 0.28)	(0.62, 0.09, 0.29)	(0.62, 0.09, 0.30)	-0.008	0.017
	ONLINE-B	(0.55, 0.20, 0.25)	(0.58, 0.19, 0.23)	(0.58, 0.19, 0.22)	0.013	-0.028
	ONLINE-G	(0.44, 0.15, 0.41)	(0.45, 0.14, 0.41)	(0.47, 0.16, 0.37)	0.006	-0.017
	TranssionMT	(0.55, 0.20, 0.25)	(0.58, 0.20, 0.22)	(0.58, 0.20, 0.22)	0.019	-0.036
	Unbabel-Tower	(0.54, 0.22, 0.24)	(0.61, 0.16, 0.23)	(0.42, 0.33, 0.25)	0.087	0.003
CS	Aya23	(0.49, 0.33, 0.18)	(0.64, 0.17, 0.19)	(0.33, 0.50, 0.18)	0.160	0.005
	CUNI-Doc	(0.51, 0.29, 0.20)	(0.54, 0.27, 0.19)	(0.39, 0.43, 0.19)	0.084	-0.014
	CUNI-GA	(0.48, 0.34, 0.18)	(0.49, 0.34, 0.17)	(0.42, 0.43, 0.15)	0.053	-0.020
	CUNI-MH	(0.05, 0.76, 0.19)	(0.44, 0.32, 0.24)	(0.05, 0.75, 0.21)	0.189	0.031
	Claude-3.5	(0.64, 0.18, 0.17)	(0.78, 0.06, 0.16)	(0.48, 0.35, 0.17)	0.147	-0.005
	CommandR+	(0.47, 0.36, 0.17)	(0.66, 0.12, 0.22)	(0.34, 0.48, 0.18)	0.159	0.024
	GPT-4	(0.82, 0.05, 0.13)	(0.85, 0.02, 0.14)	(0.73, 0.15, 0.13)	0.061	0.001
	IKUN	(0.53, 0.27, 0.20)	(0.64, 0.15, 0.22)	(0.54, 0.27, 0.19)	0.053	0.003
	IOL-Research	(0.74, 0.11, 0.15)	(0.78, 0.07, 0.15)	(0.69, 0.17, 0.15)	0.048	-0.001
	Llama3-70B	(0.73, 0.04, 0.23)	(0.75, 0.04, 0.22)	(0.70, 0.10, 0.21)	0.037	-0.019
	NVIDIA-NeMo	(0.57, 0.22, 0.21)	(0.61, 0.23, 0.16)	(0.59, 0.25, 0.16)	0.035	-0.048
	ONLINE-A	(0.14, 0.72, 0.14)	(0.01, 0.85, 0.13)	(0.01, 0.87, 0.12)	0.011	-0.012
	ONLINE-B	(0.82, 0.02, 0.16)	(0.78, 0.06, 0.16)	(0.76, 0.07, 0.17)	0.007	0.003
	ONLINE-G	(0.77, 0.07, 0.17)	(0.76, 0.07, 0.17)	(0.78, 0.05, 0.17)	-0.010	0.004
	ONLINE-W	(0.48, 0.37, 0.15)	(0.47, 0.37, 0.16)	(0.12, 0.72, 0.17)	0.170	0.009
ES	SCIR-MT	(0.61, 0.18, 0.21)	(0.65, 0.14, 0.21)	(0.50, 0.31, 0.19)	0.081	-0.007
	TranssionMT	(0.17, 0.69, 0.14)	(0.20, 0.67, 0.13)	(0.08, 0.80, 0.12)	0.074	-0.016
	Unbabel-Tower	(0.45, 0.35, 0.20)	(0.58, 0.19, 0.22)	(0.38, 0.42, 0.20)	0.100	0.017
	Aya23	(0.46, 0.24, 0.30)	(0.56, 0.11, 0.33)	(0.35, 0.35, 0.30)	0.107	0.010
	Claude-3.5	(0.56, 0.12, 0.32)	(0.64, 0.03, 0.33)	(0.52, 0.16, 0.32)	0.055	0.010
	CommandR+	(0.40, 0.30, 0.30)	(0.53, 0.17, 0.30)	(0.31, 0.40, 0.29)	0.112	-0.005
	Dubformer	(0.58, 0.13, 0.28)	(0.62, 0.08, 0.30)	(0.58, 0.11, 0.31)	0.006	0.018
	GPT-4	(0.68, 0.02, 0.29)	(0.68, 0.02, 0.30)	(0.66, 0.05, 0.29)	0.013	0.003
	IKUN	(0.27, 0.41, 0.32)	(0.34, 0.35, 0.31)	(0.26, 0.41, 0.33)	0.036	0.003
	IOL-Research	(0.59, 0.10, 0.31)	(0.65, 0.05, 0.30)	(0.58, 0.14, 0.29)	0.049	-0.018
	Llama3-70B	(0.64, 0.04, 0.33)	(0.69, 0.01, 0.30)	(0.59, 0.11, 0.30)	0.064	-0.024
	MSLC	(0.53, 0.09, 0.39)	(0.55, 0.08, 0.38)	(0.53, 0.10, 0.38)	0.018	-0.012
	ONLINE-A	(0.55, 0.09, 0.35)	(0.55, 0.09, 0.36)	(0.52, 0.12, 0.36)	0.009	0.007
	ONLINE-B	(0.59, 0.08, 0.34)	(0.59, 0.07, 0.34)	(0.57, 0.09, 0.34)	0.008	0.001
	ONLINE-G	(0.51, 0.15, 0.34)	(0.50, 0.16, 0.35)	(0.50, 0.15, 0.35)	-0.003	0.006
	ONLINE-W	(0.48, 0.16, 0.36)	(0.51, 0.12, 0.36)	(0.28, 0.36, 0.36)	0.116	0.001
	TranssionMT	(0.56, 0.10, 0.34)	(0.55, 0.10, 0.35)	(0.52, 0.13, 0.34)	0.011	0.010
	Unbabel-Tower	(0.42, 0.22, 0.36)	(0.47, 0.16, 0.37)	(0.34, 0.32, 0.33)	0.077	-0.011

Table 7: Effect of binary **gender stereotypes** on the proportion of masculine (M), feminine (F), and gender-neutral translations (N). Subscripts denote the stereotype influencing the assumed gender of the adjective referent. The average effect on binary gender is denoted by ΔG_{avg} , and the average effect on the proportion of neutral translations is denoted by ΔN_{avg} .