GATE X-E : A Challenge Set for Gender-Fair Translations from Weakly-Gendered Languages

Spencer Rarrick **Ranjita Naik**[*†] **Sundar Poudel** **Vishal Chowdhary**

Abstract

Neural Machine Translation (NMT) continues to improve in quality and adoption, yet the inadvertent perpetuation of gender bias remains a significant concern. Despite numerous studies on gender bias in translations into English from weakly gendered-languages, there are no benchmarks for evaluating this phenomenon or for assessing mitigation strategies. To address this gap, we introduce GATE X-E, an extension to the GATE (Rarrick et al., 2023) corpus, that consists of human translations from Turkish, Hungarian, Finnish, and Persian into English. Each translation is accompanied by feminine, masculine, and neutral variants. The dataset, which contains between 1250 and 1850 instances for each of the four language pairs, features natural sentences with a wide range of sentence lengths and domains, challenging translation rewriters on various linguistic phenomena. Additionally, we present a translation gender rewriting solution built with GPT-4 and use GATE X-E to evaluate it. We open source[1] our contributions to encourage further research on gender debiasing.

1 Introduction

Despite dramatic improvement in general NMT quality and breadth of supported languages over recent years (Team et al., 2022), gender bias in NMT output remains a significant problem (Piazzolla et al., 2023). One such type of gender bias is spurious gender-markings in NMT output when none were present in the source. This occurs most frequently when translating from a weakly-gendered language into a more strongly gendered one. We explore this phenomenon in translations from Turkish, Persian, Finnish, and Hungarian into English.

Gender can be marked in English through gendered pronouns (*he*, *she*, etc.) and possessive determiners (*his*, *her*), or through a limited number of intrinsically gendered nouns (*mother*, *uncle*, *widow*, etc), many of which are kinship terms.

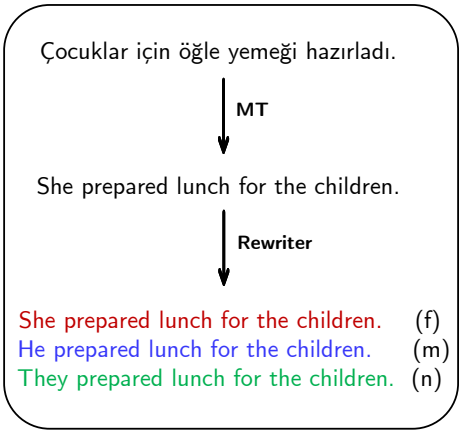

Figure 1: **Gender Bias in Turkish-English Translation.** When translating from Turkish to English, the model tends to use the female pronoun *she* for gender-unspecified individuals, likely due to a perceived link between women and child care. This bias can be mitigated by providing feminine, masculine, and neutral rewrites.

In each of our selected set of source languages, all personal pronouns are gender-neutral, such as *O* in Turkish meaning *he/she/singular they*. These languages do use some intrinsically gendered noun words, but not necessarily for all of the same concepts that English does. Turkish differentiates *mother* (anne) from *father* (baba), but does not differentiate *nephew* from *niece* (both are *yeğen*).

This difference in gender on third-person singular pronouns leads to translation scenarios such as the one seen in Figure 1, where someone with no specified gender in the source is marked as female in the translation through the pronoun *she*. NMT models often make gender assignments according to stereotypes (Stanovsky et al., 2019; Costa-jussà et al., 2023) - in this case a model appears to associate child care with women. One remedy for this

[*]All authors are affiliated with Microsoft.
[†]Contact author at ranjitan@microsoft.com.
[1]https://github.com/MicrosoftTranslator/GATE-XE

Findings of the Association for Computational Linguistics ACL 2024, pages 8526–8546
August 11-16, 2024

category of problems is to supplement the default feminine translation with masculine and gender-neutral alternatives, so that all possible gender interpretations are covered. This can be accomplished by applying a gender rewriter to the original NMT output, as shown in the bottom portion of Figure 1.

GATE (Gender-Ambiguous Translation Examples, Rarrick et al. 2023) introduced an evaluation benchmark for gender rewrites for translations from English into French, Spanish, and Italian. In this work, we introduce GATE X-E[2], an extension to GATE that focuses on translations *into* English from a set of more weakly-gendered languages. It consists of natural sentences with strong diversity of sentence lengths and domains, and challenges translation rewriters on a wide range of linguistic phenomena. GATE X-E contains between 1250 and 1850 instances for each of our language pairs.

We also present a translation-rewriting solution that utilizes GPT-4 (OpenAI, 2022) to provide gendered and gender-neutral alternatives. It achieves high accuracy on the pronoun-only subset of GATE X-E. Finally, we also perform human evaluation and provide a detailed error analysis of the results.

The remainder of this paper is organized as follows – In section 2, we discuss the corpus creation process and structure of GATE X-E. In section 3, we discuss how various properties can affect the difficulty of translation rewriting problems. In section 4 we introduce a GPT-4-based translation-rewriting solution. In section 5 we discuss results of our experiments and perform detailed error analysis. Finally, In section 6 we cover related work.

2 GATE X-E Dataset

We introduce GATE X-E by describing the annotation process and labels used, as well as providing statistics on the collected data.

2.1 Arbitrarily Gender-Marked Entities

Following Rarrick et al. (2023), we use Arbitrarily Gender-Marked Entity (AGME) to refer to individuals whose gender is not marked in a source sentence, but is in a translation, either through a gendered pronoun or an intrinsically gendered noun. Presence of an AGME in a translation indicates that alternative gender translations are possible.

The subject pronoun from the example translation shown in Figure 1 is an AGME. Because there is no gender marking in the source sentence, it is valid to translate the subject as *she*, *he* or *they*.

2.2 Dataset Creation and Annotation

All instances in GATE X-E consist of a single source sentence with one or more translations covering possible gender interpretations. We pulled sentence pairs for each of our language pairs from several corpora found on OPUS[3]: Europarl (Koehn, 2005), TED talks (Raine, 2020), tatoeba[4], wikimatrix (Schwenk et al., 2021), OpenSubtitles (Lison and Tiedemann, 2016), QED (Lamm et al., 2021) and CCAligned (El-Kishky et al., 2020). We then apply the following filters:

- The source sentence scores at least 0.7 match for the intended language when using the python langdetect[5] package.
- The English translation contains at least one word on a curated word list consisting of 79 English nouns (e.g. *mother, uncle, actress, duke*) and pronouns (*he, she, him, her, his, hers, himself, and herself.*) This list is found in Table 13 in the appendix.

We then sampled sentences from the filtered set and provide them to annotators. From this data, the annotators selected appropriate sentences and annotated them for entity types, number of AGMEs, and gendered-alternative translations if AGMEs are present. To be included, a translation must include at least one gender-marked term in the target, which could be a pronoun[6] or noun[7].

For each language, a second annotator then reviewed the data to correct errors and inconsistencies. Across pairs language pairs, the second annotator agreed with the first annotator's assessment 95% of the time. In the remaining 5% of cases a consensus was reached after discussion between the two annotators. All of the annotators are native speakers of the source language, fluent in English, and hold advanced degrees in linguistics or a related field.

If there are one or two AGMEs in the translated pair, they will provide translation variants so that all possible gender combinations for those AGMEs (among female, male, and neutral) are covered.

[2]X-E indicates translation from 'X' language into English

[3]https://opus.nlpl.eu/

[4]https://tatoeba.org/en/about

[5]https://pypi.org/project/langdetect/

[6]including possessive determiners *his*, *her*, *their*

[7]including other gender-marking modifiers, such as *female*

They do so by replacing all gendered pronoun and noun mentions with corresponding words of the respective gender. Neutral variants are omitted if there is no suitable gender neutral term in common usage for a concept. For example, the term *nibling* exists as a gender neutral variant of *niece* or *nephew*, but is not in common usage and so neutral variants of translations using *niece* or *nephew* would be left out.

Some sentences may contain a mixture of references to AGMEs as well as to humans who are gender-marked in the source. In these cases, gender indicated in the source will be preserved in all translations, as in *father* and *his will* in the example shown in Figure 2. In this example, *Babası* explicitly indicates *father* in the source.

Src	Babası vasiyetinde arabayı ona bıraktı.
Fem	Her father left her the car in his will.
Masc	His father left him the car in his will.
Neut	Their father left them the car in his will.
Lbls	target_only_gendered_pronoun, source+target_gendered_noun+pronoun, 1-AGME, mixed

Figure 2: **GATE X-E Example Instance.** This includes Turkish source; feminine, masculine and gender-neutral English translations; and labels.

2.3 Labels

The labels used in GATE X-E are defined in Table 1, along with examples for each. All instances in GATE X-E refer to at least one person who is marked for gender in the English target. We include both positive and negative examples. In positive examples, at least one of those individuals was not marked for gender in the source, and is therefore an AGME, meaning that alternative translations with different gender markings are possible. In negative examples, all individuals who are marked for gender in the target are also marked in the source, so no alternative translations are possible.

Each instance will typically have multiple labels. There will always be one label indicating AGME count and each will have an associated label beginning with `target_only_gendered_*`, though each distinct label will be included only once. Any number of negative labels beginning with `source_*` may appear with any instance, as the presence of a non-AGME in a sentence does not affect AGME count.

`name` will be included if an AGME is referred to by name, but there will always be another positive label as well. `non-AGME-name` indicates that a name appeared in the instance that was not associated with an AGME. Instances marked `mixed` mention at least one AGME and one individual who is gender-marked in the source.

2.4 Corpus Statistics

Table 4 in the appendix provides a comprehensive breakdown of corpus statistics for GATE X-E, with instance counts per language for each label.

More than half of the instances for each language pair have exactly one AGME, with around 20-30% being negative instances, having no AGMEs at all. Most AGMEs are `target_only_gendered_pronoun`, meaning that they have no gender markings in the source and the only words in the target that mark their gender are pronouns. This is in part because there are relatively few nouns which are gendered in English but not gendered in the source languages.

Non-AGME references involve a gendered noun in the source, and for most of the languages about half of these also include a pronoun reference.

Each language pair contains around 250 instances labeled `mixed`. Between 15% and 25% of instances per language have the `name` label, while less than 10% per language have `non-AGME-name`. We present the distribution of sentence lengths in source and target languages in Figure 3.

3 Translation Gender Rewriting

Translation gender rewriting is the process of taking a translated source-target pair and producing alternative translations with different gender markings. In a correctly rewritten translation, the gender markings should remain compatible with all gender information found in the source sentence (Habash et al., 2019). We consider this problem from the viewpoint of a user who wishes to see a set of three gendered-alternative translations with uniform output gender side-by-side: all-female, all-male and all-neutral. Because the translations will be viewed as a set, the translations should only vary from one-another in specific words that mark gender.

Here we discuss the difference in difficulty between rewrite problems where gendered nouns are included and those where gender is only marked by pronouns.

Description	Example (tr > en)
Negative/Non-AGME labels	
source+target_gendered_noun A person is referred to by a gendered noun in both source and translations.	Git ve **erkek kardeş**ine yardım et. → Go and help your **brother**.
source+target_gendered_noun+pronoun A person referred to by a gendered noun in the source is referred to by both a gendered noun and one or more gendered pronouns in the translations.	**Anne**m zaten kararını verdi. → My **mom** has already made **her** decision.
source_gendered_noun_target_pronoun A person is referred to by a gendered noun in the source, and one or more gendered pronouns in the translations (but not by a gendered noun).	**O**, gerçek bir **bilim adamı**dır. → **He** is a **scholar** to the core. (*bilim adamı* indicates a male scholar)
non-AGME-name A non-AGME person is referred to by name.	**Umut**'un *torunu* ünlü bir yazar değil mi? → **Umut**'s *granddaughter/grandson/grandchild* is a famous writer, isn't/aren't *she/he/they*?
Positive/AGME labels	
target_only_gendered_noun A person who is not gender-marked in the source is referred to with a gendered noun in the translations.	**Yeğen**im bugün geliyor. → My **niece/nephew** is coming today.
target_only_gendered_pronoun A person who is not gender-marked in the source is referred to with a gendered pronoun in the translations.	**Onun** yardımı paha biçilmezdi. → **Her/His/Their** help has been invaluable.
target_only_gendered_noun+pronoun A person who is not gender-marked in the source is referred to with both a gendered noun and gendered pronoun in the translations.	**Torun**un iş**ini** seviyor olmalı. → Your **granddaughter/grandson/grandchild** must love **her/his/their** job.
name An AGME is referred to by name. We treat personal names as non-gender-marking.	**Beyza** akşam yemeğini bitiremedi. → **Beyza** wasn't able to finish **her/his/their** dinner. (*Beyza* is typically considered a feminine name)
Other	
mixed Both positive and negative examples are present	*Baba***sı** yine uçağ*ını* kaçırdı. → **Her/His/Their** *father* missed *his* plane again.
***N* AGME(s)** *N* is a whole number representing the number of AGMEs in the instance. Negative examples are annotated as 0 AGMEs.	0 AGME: My mother read her book. 1 AGME: **She/He** ate **her/his** lunch alone. 2 AGME: **She/He** annoyed *her/him* with **her/his** music.

Table 1: **Label Definitions and Examples.** Words relevant to the label are bolded or italicized in source and target. *Pronoun* in these definitions includes possessive determiners *her*, *his*, *their*.

Pronoun-Only Problems

For our source languages, if the only gender markers in the target sentence are gendered pronouns, there typically cannot be gender markers in the source sentence, since those languages do not have any gendered pronouns. We can therefore assume that if we have no gender information from external context, then all individuals mentioned by a gendered pronoun in the translation must be AGMEs.

Rewriting in this scenario reduces to the relatively simple task of adjusting surface forms of all pronouns to match the desired gender. For rewrites involving *he* or *she* to gender-neutral *they*, some verbs must additionally be adjusted to compatible surface forms. Since our focus is on rewrites with uniform gender assignments in the output (all-female, all-male, or all-neutral), this removes any need to determine which pronouns refer to which individual where more than one is mentioned.

Here we see an example of a Pronoun-only instance with two AGMEs:

Female + Female	She gave her her umbrella.
Male + Male	He gave him his umbrella.
Neutral + Neutral	They gave them their umbrella.

Instances of GATE X-E that fall into this subtype will always have the label `target_gendered_pronoun_only`, and never any labels with containing `gendered_noun`.

Note that there is an exceptional scenario, where external context makes it reasonable for a gendered noun in one of our source languages to be translated into a gendered pronoun in English. See the example for `source_gendered_noun_target_pronoun` in Table 1, and refer to Appendix 10 for further discussion.

Gendered-Noun Problems

If we expand our scope to include translations containing gendered nouns, we encounter several new challenges that render the rewriting problem significantly more difficult. The following pair of examples illustates some of those new challenges.

Kardeşine ziyarete gelip gelmeyeceğini sordu.

⇓

He asked his sister if she would visit.

In this translation, both the male and female individuals are AGMEs since *Kardeşine* simply denotes a *sibling* without any gender specification. Therefore, there are nine possible rewrites, including the original translation: *He* and *his* can be optionally replaced with *she/her* or *they/their*, and *sister/she* can be optionally replaced with *brother/he* or *sibling/they*.

In the next example, however, the gender of the sibling is specified in the source as female by the addition of the word *Kız*, even though the default English translation is exactly the same:

***Kız** kardeşine ziyarete gelip gelmeyeceğini sordu.*

⇓

He asked his sister if she would visit.

Here *sister* must remain fixed because of the gender marking in the source. *She* is also fixed because it is coreferent with the sister. Only the individual referred to by *he* and *his* is an AGME, so only three valid rewrites exist (including the original):

Female	She asked her sister if she would visit.
Male	He asked his sister if she would visit.
Neutral	They asked their sister if she would visit.

More specifically, the additional challenges inherent in this problem class include the following.

- Gender-marked nouns may appear on the source as well, so we must examine both source and target to determine if variants are needed. This is demonstrated by the behavior of *sister* in the two examples above.

- Gendered pronouns in the target may refer to individuals whose gender was marked in the source, and are therefore not appropriate to modify. When multiple individuals are mentioned, we must differentiate which ones refer to such non-AGME individuals to produce a correct rewrite. This is demonstrated by the behavior of *she* in the two examples above.

A system that is capable of solving these problems must then be able to implicitly perform coreference resolution and alignment of nouns between the source and target, significantly increasing the complexity of a solution.

4 Experiments

We leverage GPT-4 to implement a translation gender-rewriting solution and evaluate it on GATE X-E.

4.1 Rewriting with GPT-4

Our solution uses chain-of-thought prompting (Wang et al., 2023) to elicit GPT-4 to produce three variant translations for each input source-translation pair – all neutral, all feminine and all masculine, while leaving any gendered words associated with non-AGMEs unmodified. If all AGMEs had been marked with the same gender in the input translation (all masculine or all feminine), the corresponding output is expected to be identical to the input translation.

Each step in the prompt is accompanied by detailed clarifications and example vocabulary. The prompt also includes three full examples, customized per source language. The examples indicate that "None" should be returned in lieu of translation variants when there are no AGMEs present. The full prompt for Turkish-English can be found in the appendix in Figures 7 and 8.

In order to iterate and identify a suitable prompt for use in the rewriting task, we used a small dummy dataset of 100 English-Turkish sentence pairs, distinct from those in GATE X-E. During early experiments with simpler prompts, we found that GPT-4 would often make incorrect assumptions about what individuals has gender markings in the source or input translation. We found that making instructions much more explicit helped reduce the frequency of these assumptions.

4.2 Data Preparation

Each GATE X-E instance consists of a source sentence and a set of translations in English. Each positive instance is used multiple times during evaluation. More formally, we transform each instance into multiple *test tuples*, each corresponding to a different evaluation setting.

Each test tuple consists of a *source sentence*, an *original translation* and a *rewriter reference output*. In all cases we use one of the instance's reference translations as the test tuple's original translation, and another reference translation from the same instance as the rewriter reference output. All-feminine and all-masculine original translations are paired with opposite-gender and all-neutral rewriter reference outputs. In instances with two AGMEs, we also evaluate each non-neutral mixed-gender reference as an original translation (i.e. feminine/masculine and masculine/feminine) and pair with all-feminine, all-masculine and all-neutral rewriter reference outputs.

Negative instances, which do not include AGMEs and should not be modified by the rewriter, are handled separately. For each negative instance we create tuples to check that the rewriter's feminine, masculine and neutral outputs are all unmodified from the original translation.

At the top level, we group the above-described test tuples by the gender-assignments of AGMEs in their original translation, producing four subsets: feminine, masculine, mixed-feminine-and-masculine (with 2 AGME), and negative. These correspond to major sub-categories in our evaluation results.

For simplicity, we exclude the handful of instances with three or more AGMEs. We do not include any tuples that include gender-neutral forms in the original target because of the ambiguity in distinguishing singular neutral pronouns from plural *they* forms. For any instance where the neutral reference is empty (e.g. because there is no neutral form of a term), no test tuple with neutral rewriter reference output is created.

Each test tuple is also marked with a subtype. If it came from an instance containing any intrinsically gendered nouns in either source or target, it is designated *Gendered-Noun*. Positive tuples from instances without gendered nouns must have target-side gendered pronouns as their only gender markings and are designated *Pronoun-Only*. Note that negative examples always contain a gendered noun in the source, and so they do not have a distinction between the two types.

Table 5 in the appendix shows counts for each source language and category. Note that each tuple with an all-feminine original translation has a corresponding one that is all masculine, so the counts are identical for these sets, marked (f/m) in the table.

5 Results

Table 2 shows the accuracy of our solution on test tuples over each combination of source language, and subtype, with *Overall* indicating an aggregate score over *Pronoun-Only* and *Gendered-Noun*. The top labels in the header row indicate gender of AGMEs in the original target, and the bottom indi-

Language	Subtype	Fem Orig ↑		Masc Orig ↑		Mixed Orig ↑			Negative ↑	
		M	Neut	F	Neut	F	M	Neut	Gender	Neut
tu → en	Overall	0.81	0.86	0.80	0.85	0.78	0.82	0.57	0.87	0.80
	Pronoun-Only	0.99	0.99	0.96	0.99	0.93	0.96	0.98	-	-
	Gendered-Noun	0.59	0.63	0.60	0.63	0.75	0.80	0.45	-	-
fi → en	Overall	0.90	0.80	0.82	0.81	0.67	0.83	0.63	0.89	0.88
	Pronoun-Only	0.98	0.98	0.97	0.98	0.94	0.96	0.97	-	-
	Gendered-Noun	0.75	0.53	0.54	0.57	0.34	0.68	0.54	-	-
hu → en	Overall	0.85	0.79	0.84	0.81	0.75	0.92	0.78	0.84	0.84
	Pronoun-Only	0.98	0.99	0.96	0.98	0.92	0.97	0.97	-	-
	Gendered-Noun	0.66	0.61	0.68	0.76	0.56	0.86	0.56	-	-
fa → en	Overall	0.86	0.81	0.86	0.81	0.81	0.84	0.66	0.54	0.53
	Pronoun-Only	0.99	0.98	0.99	0.98	0.93	0.93	0.96	-	-
	Gendered-Noun	0.67	0.54	0.68	0.55	0.78	0.81	0.58	-	-

Table 2: **Accuracy of our Rewriting Solution.** Accuracy on test elements for each source language, problem subtype, original target gender (top header row), and requested output gender (second header row). Only exact matches to reference are counted.

cates the desired output gender.

Following Rarrick et al. (2023), we focus on exact match accuracy to the reference. Frequently only one or two words will be different between an original translation and a correct rewrite. In this context, metrics such as BLEU (Papineni et al., 2002) and WER are not very effective at determining the significance of single extraneous or missed word modifications. Therefore, credit is only given for a test tuple when the rewriter output exactly matches the reference output.

5.1 Pronoun-Only Subset

On the Pronoun-Only subset, the solution rarely makes mistakes for masculine and feminine original targets, with scores ranging from 0.96 to 0.99. Test cases where the original target has mixed gender all come from 2-AGME instances. These skew towards longer and more complicated sentence, which thus leads to slightly lower accuracy.

On most language pairs we see that Pronoun-Only rewrites into masculine outperform rewrites into feminine by a few percentage points. The largest gap is 5 points for mixed-gender original target on Hungarian. This may indicate a slight general tendency of GPT-4 to prefer phrasing using masculine pronouns.

5.2 Gendered-Noun Subset

Scores on the Gendered-Noun subset are substantially lower than for *Pronoun-Only*, generally ranging from about 0.5 to 0.8, with Finnish mixed→feminine as an outlier at the low end at 0.34. The score differential with *Pronoun-Only* can be mostly attributed to the more difficult nature of the problems. However, there are some cases where there are multiple acceptable alternative phrases to use in a rewrite, and GPT-4 chooses a different one from the reference.

This effect is most pronounced in the Finnish *Complex* mixed-original target data. This subset contains a large amount of data from Europarl that includes titles such as *Mr.* and *Mrs.* and addresses to *Mr. President*. The feminine rewrites often choose a mismatched form, such as *Ms. Müller* rather than *Mrs. Müller*, or *Mrs. President* rather than *Madam President*.

Similarly, neutral rewrites on such sentences often produces *Honorable President*, *Chairperson*, or *Honorable Speaker* as a rewrite of *Mr. President* or *Madam President*, mismatching *Honored President* in the reference.

5.3 Negative Subset

Negative→Gender score indicates how often both the feminine and masculine outputs produced by our solution exactly matched the original translation, while *Negative→Neutral* measures the same for the neutral output. An output of *None* from GPT-4 is also possible, indicating that all variants should be considered a copy of the original translation. On the negative data subset, this is considered a reference match.

For Turkish, Finnish and Hungarian, we see scores for both *Gender* and *Neutral* subsets in the 0.8 to 0.9 range. Farsi is an outlier with 0.54 and 0.53 at the low end.

With the exception of Turkish, we find that we almost never see matches on negative test items aside from *None* outputs. For Turkish, 47% of

Error Type	Pronoun-Only ↑		Gendered-Noun ↑		Negative ↑	
	Gen (%)	Neut (%)	Gen (%)	Neut (%)	Gen (%)	Neut (%)
Extraneous noun change	12.5	10.0	2.8	3.5	0.0	34.5
Extraneous pronoun change	0.0	10.0	4.0	5.6	100	65.5
Missing noun change	0.0	0.0	55.5	51.9	-	-
Missing pronoun change	87.5	80.0	37.5	38.6	-	-
Total errors	30	8	427	381	57	58

Table 3: **Distribution of Errors from Human Evaluation for Turkish-English.** Shows percentages of errors over Pronoun-Only, Gendered-Noun and Negative subsets of the data, for gendered and neutral requested outputs.

non-*None* outputs match for gendered and 15% of neutral outputs. For all languages, however, we do see a large number of neutral outputs that match a version of the original translation where all pronouns are modified to their neutral variants[8]. This occurs in particular when there are gendered nouns in the source which determine the gender of some pronouns in the translation as well. For example, instead of *the man has something under his coat*, it would output *the man has something under* ***their*** *coat*.

If were relax matching criteria to allow this variant, neutral negative accuracy increases to 0.91 for Farsi, 0.92 for Turkish, 0.95 for Finnish and 0.95 for Hungarian.

5.4 Gender-Neutral Rewriting

For most settings, we observe that accuracy is similar whether rewriting into gendered or gender-neutral output. The most obvious outlier is that for *Gendered-Noun* mixed-gender original translations, neutral output scores are consistently significantly lower than the masculine output scores. Some of this gap can be attributed to mismatches in noun forms, such as *Honorable Speaker* and *Honored Speaker* as mentioned above.

We also see a few other data points where neutral outputs significantly trail their gendered counterparts on *Gendered-Noun* sets: Farsi feminine and masculine original targets, and Finnish feminine original target.

5.5 Human Evaluation

For the Turkish data, one of the Turkish annotators provided annotations for error type on all outputs that did not exactly match the reference. These are found in Table 3.

[8] and verb agreement modified to match

For each test item, they mark whether there nouns or pronouns were changed where the reference and original translation matched (*extraneous noun/pronoun change*, as well as whether any nouns or pronouns should have been changed but were not (*missing noun/pronoun change*). Distributions for masculine and feminine outputs were similar, so we show combined *gendered* outputs for each subtype. We also aggregate over mixed- and uniform-gender inputs.

For positive test cases, missing noun and pronoun changes were far more common than extraneous changes. For cases containing gendered nouns, noun changes were missed more often than pronoun changes, while extraneous pronoun changes were more common than extraneous noun changes.

For Negative test cases, gendered output only ever contained extraneous pronouns changes, while neutral outputs did have a fair number of extraneous noun changes. An example of these is changing *man* to *person* even when *man* was gender-marked in the source sentence.

Among missing pronoun errors, missing possessive determiners was by far the most common, with subject and object pronoun errors roughly equivalent. Missed and extraneous reflexives were extremely rare. We also saw a single case of subject-verb agreement error each when changing *he* and *she* to *they*, and one case where other wording was incorrectly changed in the sentence.

6 Related Work

Understanding and Assessing Gender Bias: It has been documented that MT systems often make mistakes and show gender biases when translating between languages with differing gender norms (Stanovsky et al., 2019; Prates et al., 2019; Rescigno et al., 2020; Lopez-Medel, 2021; Prates

et al., 2019; Fitria, 2021; Saunders and Byrne, 2020). Costa-jussà et al. (2023) show an 8-bleu-point gap between masculine and feminine references when translating from non-gender-marked English sentences into various strongly-gendered langauges.

Měchura (2022) proposes a taxonomy for describing situations where properties such as gender, number or formality are unknown in a source sentence but assumed in a translation, resulting in bias.

Evaluation Benchmarks:
Translating from English: Bentivogli et al. (2020) and Savoldi et al. (2022) introduced the MuST-SHE dataset, which includes triplets of audio, transcript, and reference translations for English to Spanish, French, and Italian languages, classified by gender. Stanovsky et al. (2019) developed the WinoMT challenge set, which includes English sentences with two animate nouns, one of which is coreferent with a gendered pronoun.

Renduchintala et al. (2021) introduced the SimpleGEN dataset for English-Spanish and English-German language pairs, which includes short sentences with occupation nouns and clear gender indications. The Translated Wikipedia Biographies dataset includes human translations of Wikipedia biographies for gender disambiguation evaluation. Lastly, Currey et al. (2022) presented the MT-GenEval dataset, which includes gender-balanced, counterfactual data in eight language pairs, specifically focusing on translation from English into eight widely-spoken languages.

Translating into English: Numerous studies have focused on evaluating bias in translating from from a weakly gendered language such as Turkish into English. (Prates et al., 2019; Fitria, 2021; Ciora et al., 2021; Ghosh and Caliskan, 2023).

Strategies for Gender Bias Mitigation : Habash et al. (2019) propose a gender-awareness wrapper for Arabic MT systems and develop a corpus for first-person-singular gender identification and reinflection. Alhafni et al. (2020) present an Arabic sentence-level gender reinflection approach using linguistically enhanced sequence-to-sequence models. Alhafni et al. (2022) define gender rewriting in Arabic contexts involving two users and develop a multi-step system combining rule-based and neural rewriting models.

To mitigate gender bias when translating queries that are gender-neutral in the source language, Google Translate announced a feature (Kuczmarski, 2018; Johnson, 2020) that provides gender-specific translations. Both Sun et al. (2021) and Vanmassenhove et al. (2021) have explored monolingual gender-neutral rewriting of English demonstrating that a neural model can perform this task with reasonable accuracy. Ghosh and Caliskan (2023) evaluate gender bias in GPT-3.5 Turbo output when translating from gender-neutral languages into English. To the best of our knowledge, our work is the first to leverage GPT-4 for mitigation.

Saunders and Byrne (2020) manually construct a gender-balanced profession dataset and use it to fine tune MT models to improve gender accuracy when source gender is unambiguous. Piergentili et al. (2023) provide an in depth discussion of the challenges and intricacies of generating gender-neutral translation output, focusing on English-Italian translation.

7 Conclusion

We have presented GATE X-E, a diverse dataset covering a wide range of scenarios relevant to translation gender-rewriting for English-target language pairs, covering gendered and gender-neutral rewrites. We have discussed intricacies of English-target translation rewriting, and explained what properties lead to easier or more difficult rewrite problems. We have explored the ability of GPT-4 to provide rewrites for these translations, showing that it can achieve very high accuracy on pronoun-only rewriting problems, but performs less well when gendered nouns are introduced.

In the future we hope to expand the set of source languages beyond the four weakly gendered languages used in this work, to include those with different gender nuances. As can be seen from GPT-4's inconsistent performance in translation gender-rewriting, particularly when gendered nouns are introduced, additional methods need to be explored in order to achieve sufficiently accurate and unbiased translation alternative sets.

We also hope that by making GATE X-E accessible to the broader research community, we can encourage further research on gender debiasing in the machine translation space.

8 Limitations

Our study has some limitations that could be addressed in future research. Firstly, while we utilized GPT-4 for rewriting tasks, the potential of open-source models remains unexplored and could be

beneficial. Secondly, our rewriter operates on few-shot chain-of-thought prompting. Future investigations could consider exploring the zero-shot setting, which could potentially be more cost-efficient.

During construction of our datasets, all four language pairs were given equal priority, and they are of roughly equal quality to the best of our knowledge. However, during error analysis, we were only able to secure sufficient budget to perform human annotation for a single language pair, English-Turkish.

9 Ethical Considerations

We acknowledge that our work is currently limited to English as the target language and four weakly gendered languages - Turkish, Hungarian, Finnish, and Persian - as the source languages. This focus may inadvertently create a bias towards these specific languages and their unique gender structures, potentially limiting the applicability of our findings to other languages with different gender nuances.

Additionally, our challenge set was constructed with the assistance of bilingual linguists for each language pair, which may introduce another layer of bias based on their individual interpretations and understanding of gender in language. While we plan to expand the scope of source languages in future work, these current limitations should be considered when interpreting our results.

Our work also explores generation of gender-neutral translation variants in English. The gender-neutral variants are limited to those following a singular *they* pattern and do not cover the full range of possible gender-neutral or non-binary pronouns.

Linguists who we recruited to carry out annotation tasks were sourced and managed through a linguistic services agency. They were paid on an hourly basis.

10 Acknowledgements

We thank Huda Khayrallah, Roman Grundkiewicz and Vikas Raunak for their valuable feedback. The study would not have been possible without the contribution of our annotators.

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2020. Gender-aware reinflection using linguistically enhanced neural models. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Chloe Ciora, Nur Iren, and Malihe Alikhani. 2021. Examining covert gender bias: A case study in turkish and english machine translation models.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

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

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Tira Nur Fitria. 2021. Gender bias in translation using google translate: Problems and solution. *Language Circle: Journal of Language and Literature*, 15(2).

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*,

pages 155–165, Florence, Italy. Association for Computational Linguistics.

Melvin Johnson. 2020. A scalable approach to reducing gender bias in google translate.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

James Kuczmarski. 2018. Reducing gender bias in google translate.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. QED: A Framework and Dataset for Explanations in Question Answering. *Transactions of the Association for Computational Linguistics*, 9:790–806.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Maria Lopez-Medel. 2021. Gender bias in machine translation: an analysis of google translate in english and spanish. *Academia Letters*.

Michal Měchura. 2022. A taxonomy of bias-causing ambiguities in machine translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173, Seattle, Washington. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

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

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2023. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. 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.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2019. Assessing gender bias in machine translation – a case study with google translate.

Paul Raine. 2020. Talk corpus: A web-based corpus of ted talks for english language teachers and learners.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 845–854, New York, NY, USA. Association for Computing Machinery.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online. Association for Computational Linguistics.

Argentina Anna Rescigno, Eva Vanmassenhove, Johanna Monti, and Andy Way. 2020. A case study of natural gender phenomena in translation a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. In *Computational Linguistics CLiC-it*, page 359.

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

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english.

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

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

A Further details on GATE X-E

Table 4 provides a comprehensive breakdown of corpus statistics for GATE X-E, with instance counts per language for each label. Please note that each instance will typically contain multiple labels. For example, a 1-AGME instance will have one *positive* label and zero or more *negative* labels, as any number of other non-AGME individuals may be referenced.

Figure 3 presents boxplots that demonstrate the sentence length distribution on source and target for four language pairs: Finnish to English, Hungarian to English, Persian to English, and Turkish to English. The left plot shows the sentence lengths in the source languages, and the right plot displays the sentence lengths in English, the target language. The legend indicates the color corresponding to each language pair. Compared to the other three language pairs, Finnish to English contains longer sentences.

Table 5 shows counts for each source language and category. Note that tuples with uniform gender original translations (f/m) are always created in pairs, so they have the same counts.

B Monolingual Rewriting with GPT-3.5 Turbo

GPT-4 performs very well on the pronoun-only subset of examples. However, its inference cost is high. Therefore, we evaluate the pronoun-only subset using GPT-3.5 Turbo.

B.1 Gender-Neutral Rewriting

As shown in 3, pronoun-only uniform-gender do not require access to source information, and so our GPT-3.5 Turbo solution is only given access to original translation target. We first use GPT-3.5 Turbo to produce an all-neutral rewrite and then use a rule based solution to convert the all-neutral rewrite to gendered rewrites.

The trickiest aspects of the gender-neutral rewrite are disambiguating pronoun classes for *her* and *him*, and adjusting verb forms when subjects change from *she/he* to *they*, so these are the primary decisions that GPT-3.5 Turbo must make.

We experiment with zero-shot and few-shot approaches. The zero-shot approach uses a single-sentence prompt as seen in Figure 5. The few-shot approach expands on this prompt by adding five examples, and it can be seen in Figure 6.

B.2 Gender-Neutral to Gendered Rewriting

A useful simplifying observation is for uniform-gender pronoun only rewrites, we can generate a correct feminine or masculine rewrite from the original target and a correct all-neutral rewrite. Referring back to 12, we see that all elements in the neutral column are unique, while the masculine and feminine columns each have one surface form fitting two categories. Knowing the correct neutral pronoun fully determines what pronouns should be used for a given gender.

In practice the neutral rewrite may contain errors. To minimize their impact on the gendered rewrites, we begin with the original translation, and map to pronouns directly to the desired gender where unambiguous. When ambiguous (i.e. for *her* or *his* in the original target), we rely on the chosen form of the neutral pronoun to disambiguate.

B.3 Experiments

B.3.1 Rewriters

Neutral rewriter Systems: We consider the following rewriting systems:

1. Rule-based system proposed by Sun et al. (2021): It uses Spacy and GPT-2 to resolve

	tr → en	**fa → en**	**fi → en**	**hu → en**
total instance count	1,429	1,259	1,832	1,308
target_only_gendered_noun	142	118	159	95
target_only_gendered_pronoun	1,074	906	1,096	914
target_only_gendered_noun+pronoun	114	49	105	115
source+target_gendered_noun	239	244	379	75
source+target_gendered_noun+pronoun	328	292	361	422
source_gendered_pronoun_target_noun	3	0	0	33
0 AGMEs	300	264	502	264
1 AGME	900	869	1,164	848
2 AGMEs	225	124	161	192
3 AGMEs	4	2	5	4
mixed	271	263	237	262
name	328	175	408	159
non-AGME-name	32	5	136	16

Table 4: **GATE X-E Statistics.** Sentence counts per language associated with each label. Each instance typically contains multiple labels.

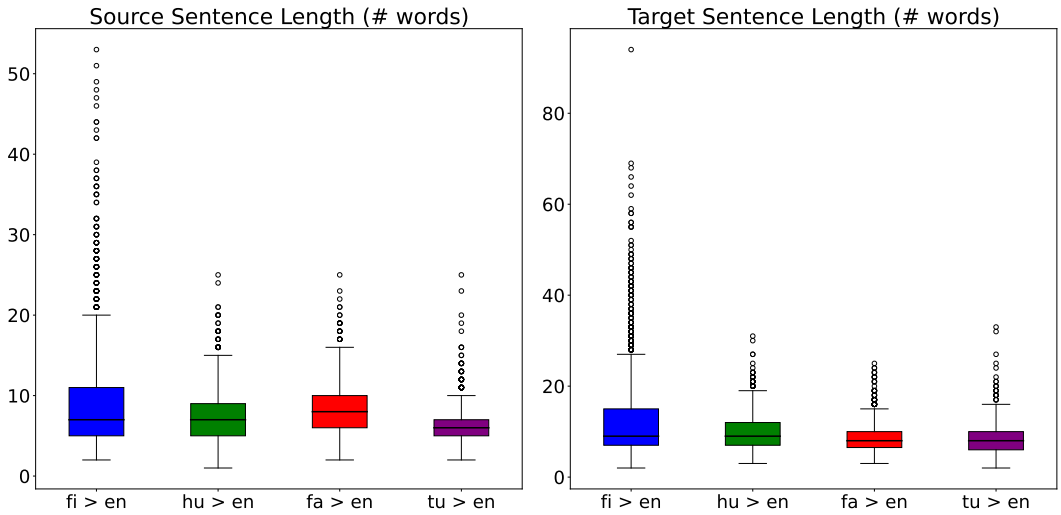

Figure 3: **Boxplots representing the distribution of sentence lengths in source and target languages.** The four language pairs are Finnish to English (fi > en), Hungarian to English (hu > en), Persian to English (fa > en), and Turkish to English (tu > en). The left plot represents the source language sentence lengths, and the right plot represents the target language (English) sentence lengths. The color of each boxplot corresponds to the language pair as indicated in the legend.

Category	tr	fa	fi	hu
Pronoun-Only (f/m)	628	857	590	580
Gendered-Noun (f/m)	500	473	454	415
Pronoun-Only (mix)	54	180	198	44
Gendered-Noun (mix)	392	142	186	200
Negative	300	502	264	264

Table 5: **Test tuple Counts By Category and Source Language.** Counts of pronoun-only, gendered-noun and negative test tuples per source language. f/m signifies count for uniform gender original targets, while mix signifies a mixture of male and female references in the original target.

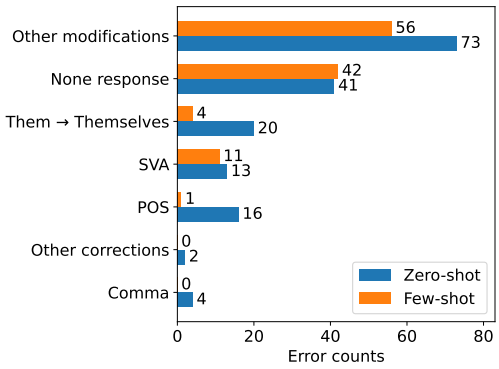

Figure 4: **Distribution of errors in GPT-3.5 Turbo's zero-shot and few-shot settings.** The majority of errors in both settings stem from unrelated modifications and the model's 'None' response, indicating no need for gender-neutral rewriting.

ambiguity with *his* and *her*, and to adjust verb forms as needed. They also trained a neural model, but it was unfortunately not accessible.

2. GPT-3.5 Turbo: We evaluate GPT-3.5 Turbo on zero-shot and few-shot settings, using the prompts shown in Figures 5 and 6 in the appendix.

We investigated the neural model introduced by Vanmassenhove et al. (2021) as well, but were unable to reproduce results on their test data.

For all GPT-based rewrites we set temperature $T = 0$

B.3.2 Evaluation

We report the rewriter systems' performance using BLEU (Papineni et al., 2002), Word Error Rate (WER), and Accuracy.

In the gender-neutral rewriting task (Table 6), GPT-3.5 Turbo performs better in the few-shot setting compared to the zero-shot setting. Although GPT-3.5 Turbo provides slightly higher accuracy compared to the rule-based system proposed by Sun et al. (2021), the rule-based system performs better based on BLEU and WER. This is because GPT-3.5 Turbo makes modifications unrelated to neutral rewriting, as detailed in the error analysis section.

In the gendered-alternatives rewriting task (Table 6), the zero-shot setting indicates that for resolving the *her→his/him* and *his→her/hers* ambiguity, gender-neutral rewrites from the zero-shot prompt are used. Similarly, the few-shot setting uses the corresponding gender-neutral outputs from the few-shot prompt. The performance of both settings is comparable.

B.3.3 Error Analysis

Figure 4 illustrates the distribution of aggregated errors across four language pairs for GPT-3.5 Turbo in both zero-shot and few-shot settings, specifically for the task of gender-neutral rewriting. The definitions of these errors are provided in Table 8 in the appendix, while Table 9 offers examples for each error label.

In both settings, the majority of errors stem from modifications unrelated to gender-neutral rewriting and from instances where the model suggests no changes are necessary to render the input text gender-neutral. Additional examples of errors due to unrelated modifications can be found in Table 10 in the appendix. The few-shot setting, however, does show an improvement in neutral rewriting errors (such as POS(part-of-speech) errors and *them* being rewritten as *themselves*) when compared to the zero-shot setting.

Tables 11 presents the error distribution for each of the four languages. Upon closer examination of the Finnish data, which has the highest error rate, we found that the errors are primarily due to the longer input length. This increases the scope for modifications of the text that are unrelated to gender-neutral rewriting.

C Prompting Templates

C.1 GPT-3.5 Turbo Prompts

Figures 5 and 6 show the GPT-3.5 Turbo zero-shot and few-shot prompts used in the gender-neutral rewriting task. We use the same prompt across all the language pairs as the task is source agnostic.

Language Pair	Method	Neutral Rewriting			Gendered Rewriting		
		Accuracy (%) ↑	BLEU ↑	WER ↓	Accuracy (%) ↑	BLEU ↑	WER ↓
tr → en	Sun et al. 2021	96.16	**99.65**	0.53	-	-	-
	Zero-shot	97.24	99.30	0.80	**99.50**	**99.90**	**0.90**
	Few-shot	**98.90**	99.55	**0.44**	99.46	99.00	0.10
hu → en	Sun et al. 2021	96.14	**99.66**	**0.53**	-	-	-
	Zero-shot	96.58	99.04	1.27	**99.27**	**99.95**	**0.08**
	Few-shot	**97.00**	99.03	1.20	99.20	99.94	0.09
fi → en	Sun et al. 2021	95.24	**99.63**	**0.62**	-	-	-
	Zero-shot	94.80	98.61	1.75	98.41	**99.85**	0.24
	Few-shot	**96.77**	98.62	1.54	**98.99**	99.80	**0.19**
fa → en	Sun et al. 2021	94.43	**99.57**	**0.65**	-	-	-
	Zero-shot	95.59	99.00	1.11	98.75	99.91	1.13
	Few-shot	**97.84**	99.16	1.01	**99.00**	**99.93**	**0.09**

Table 6: **Results of Gender Neutral and Gendered Rewriting on the Pronoun-Only Subset of GATE X-E**. We report the performance of the rule-based system proposed by Sun et al. 2021. Additionally, we evaluate GPT-3.5 Turbo in both zero-shot and few-shot settings. Gendered alternatives are generated using the algorithm described in Section B

```
Change all gendered pronouns to use singular "they" instead. Don't modify anything else : {input_text}
```

Figure 5: Zero-shot prompt template utilized in GPT-3.5 Turbo experiments.

```
Change all gendered pronouns to use singular "they" instead. Don't modify anything else.

input : His bike is better than mine.
gender neutral variant : Their bike is better than mine.

input : Jack bores me with stories about her trip.
gender neutral variant: Jack bores me with stories about their trip.

input : He kissed him goodbye and left, never to be seen again.
gender neutral variant : They kissed them goodbye and left, never to be seen again.

input : Is she your teacher?
gender neutral variant : Are they your teacher?

input : Anime director Satoshi Kon died of pancreatic cancer on August 24, 2010, shortly before her 47th birthday.
gender neutral variant : Anime director Satoshi Kon died of pancreatic cancer on August 24, 2010, shortly before their 47th birthday.

input : {input_text}
gender neutral variant :
```

Figure 6: Few-shot prompt template utilized in GPT-3.5 Turbo experiments.

I need help with a linguistic annotation task for a translation. I will give you an Turkish sentence along with its translation into English. I would like you to help me find Arbitrarily Gender-Marked Entities (AGMEs), where someone is mentioned without any marked gender in the Turkish sentence, but in the translation they have gender marking. Please follow the following steps:

1. Identify all unique individuals mentioned in the English translation in the third person and find all words that explicitly indicate those individuals' genders.
- Group words for each individual separately, considering possessive determiners (e.g., "his", "her") as referring to a separate individual from the one indicated by the noun they modify. For example, in "his uncle," "his" and "uncle" refer to two separate individuals.
- Pay attention to gender indicated by kinship terms and other gendered nouns, like "mother", "nephew", "actress".
- If the gender is explicitly indicated by pronouns in the target language, consider that gender information for the analysis. (i.e. "she", "he", "him", "her", "his", "hers", "himself", "herself" all explicitly indicate gender)
- Treat names as if they do not indicate a gender, even if they are often associated with a gender. For example, "Michael" could be either male or female, so it does not mark gender.
- Pay attention to how forms or "to be" (particularly "is") can join two mentions of the same individual. For example, in "She is my daughter," "daughter" and "she" refer to the same person.

2. Find all words in the Turkish source sentence that refer to each of the individuals found in step one.

3. For each individual, do any of the corresponding words in the Turkish source explicitly indicate a gender.
- Remember, pay attention to gender indicated by kinship words. For example, words like "erkek", "kız" , "amca", "anne" all explicitly indicate gender.
- Remember that some kinship words in Turkish are gender-neutral, such as yeğen. Do not include these as marking gender.
- Treat names as if they do not (e.g. 'Michael' can refer equally well to a man or woman).

4. Identify any instances where the gender-neutral terms in Turkish have been translated into gender-specific terms in English (AGMEs).
- Answer separately for each individual identified.

5. Next create a set of variant translations with the following notes:
- If no changes are needed, then just use the original translation exactly as it is.
- Remember to only change the words referring to AGMEs.
- if any gendered words refer to non-AGMEs, leave them untouched.
- Do not make assumptions about heterosexual relationships. Men can have husbands and boyfriends. Women can have wives and girlfriends.

Please create these three variant translations:
a. If any individuals are AGMEs and are referred to with gendered words in English, rewrite the English translation changing only those words to use their gender-neutral variants where possible. Use singular "they" instead of he, she, etc. Use "themselves" for gender neutral singular reflexives (never "themself"). Change nothing else.
b. rewrite the English translation so that any masculine words referring to AGMEs are replaced by their feminine variants. Don't change any words referring to non-AGMEs. Change nothing else.
c. rewrite the English translation so that any feminine words referring to AGMEs are replaced by their masculine variants. Don't change any words referring to non-AGMEs. Change nothing else.

Figure 7: Part I of Few-shot prompt template utilized in GPT-4 experiments.

Example 1 -
Source Sentence: Amcası kendi kendine konuşuyor.
Original Translation: His uncle talks to himself.
1. individual 1: "His" is masculine. individual 2: "uncle", "himself" are masculine.
2. individual 1: no explicit words in the source indiviudal 2: "Amcası", "kendi kendine"
3. individual 1: no words indicate gender indivudual 2: "Amcası" is masculine.
4. individual 1: AGME - masculine in translation ("His"), but gender neutral in the source (no explicit words) individual 2: not an AGME - gender is masculine in both the source ("Amcası") and translation ("uncle")
5. a. Their uncle talks to himself.
   b. Her uncle talks to himself.
   c. His uncle talks to himself.

Example 2 -
Source Sentence: Annem öğle yemeğini yalnız yiyordu.
Original Translation: My mother ate her lunch alone.
1. individual 1: "mother", "her" are feminine.
2. individual 1: "Annem"
3. individual 1: "Annem" is feminine.
4. individual 1: not an AGME since gender is feminine in both the source and translation
5. None

Example 3 -
Source Sentence: O benim kızım
Original Translation: She is my daughter.
1. individual 1: "she", "daughter" are feminine.
2. individual 1: "O", "kızım"
3. individual 1: "kızım" is feminine.
4. individual 1: not an AGME since gender is feminine in both the source and translation
5. None

Source Sentence: {source_sentence}
Original Translation: {original_translation}

Figure 8: Part II of Few-shot prompt template utilized in GPT-4 experiments.

C.2 GPT-4 Prompts

The full prompt for Turkish-English can be found in Figures 7 and 8. Prompts for other languages use the same structure, but examples are customized to fit those languages.

D Rewriting with GPT-4

Our solution uses chain-of-thought prompting (Wang et al., 2023) to elicit GPT-4 to produce three variant translations for each input source-translation pair – all-neutral, all-female and all-male, while leaving any gendered words associated with non-AGMEs unmodified. We ask it to work step-by-step through the process of identifying AGMEs before finally rewriting the original translation:

- Identify unique individuals mentioned in the target, as well as any gendered words that refer to them.
- Identify words in the source that refer to those same individuals.
- Determine which source words mark for gender.
- Designate any individuals referred to by gendered words in the target, but not in the source as AGMEs
- Produce a neutral, feminine and masculine variant translation where any gendered words referring to AGMEs are modified to match the respective gender.

Each step in the prompt is accompanied by detailed clarifications and example vocabulary. The prompt also includes three full examples, customized per source language. The examples indicate that "None" should be returned in lieu of translation variants when there are no AGMEs present. The full prompt for Turkish-English is shown found in Figures 7 and 8.

E Mitigation Strategies Based on Source

Rather than rewriting English-target translations into feminine, masculine, and neutral forms, one could use the source sentence as input to create these three variants directly. This section explains how GATE X-E can be employed to assess such a system.

The first step is to verify that the generated feminine, masculine, and neutral variants are the same, except for changes related to gender. This is a crucial step as it ensures that the meaning of the translation remains consistent, regardless of the gender. If there are differences in the translations beyond the gender-related changes, it could imply that the translation is not accurate or is introducing additional bias. After this, the generated output can be compared with the feminine, masculine, and neutral references provided in GATE X-E using contextual MT evaluation metrics.

Kuczmarski (2018) initially explored a source-based debiasing approach in which they enhanced a Neural Machine Translation (NMT) system to produce gender-specific translations. This was achieved by adding an additional input token at the beginning of the sentence to specify the required gender for translation (e.g., *<2FEMALE> O bir doktor* → *She is a doctor*). However, they encountered challenges in generating masculine and feminine translations that were exactly equivalent, with the exception of gender-related changes. As a result, they later switched to a target-based rewriting approach in their subsequent work (Johnson, 2020).

Original	Gender-Neutral	Gendered Alternatives
The teacher compared my poem with one of his.	The teacher compared my poem with one of theirs.	The teacher compared my poem with one of hers. The teacher compared my poem with one of his.

Table 7: Examples illustrating the generation of gendered alternatives using gender-neutral rewrites

Error Category	**Error Label**	**Description**
Corrections unrelated to neutral rewriting	Comma	Comma added or removed.
	Other corrections	Error corrections that are not related to making the text gender-neutral.
Neutral Rewriting	POS	Wrong form of 'they' produced by the rewriter.
	SVA	Failure to make correct subject-verb agreement.
	Them → Themselves	Second appearance of them changed to themselves.
	'None' response	The rewriter returns 'None' suggesting no changes needed to make the input gender neutral.
Irrelevant modifications	Other modifications	Modifications unrelated to neutral rewriting.

Table 8: Description of the error labels

Error Label	**Input**	**Output**	**Reference**
Comma	Well, you surprised me!**,** Afshin said as she opened the door and saw Mary standing there.	Well, you surprised me! Afshin said as they opened the door and saw Mary standing there.	Well, you surprised me!**,** Afshin said as they opened the door and saw Mary standing there.
Other corrections	I **have** never heard of him before that.	I **had** never heard of them before that.	I **have** never heard of them before that.
POS	The secretary noted down what **her** boss had said.	The secretary noted down what **they** boss had said.	The secretary noted down what **their** boss had said.
SVA	**Does** she come here every week?	**Does** they come here every week?	**Do** they come here every week?
Them → Themselves	She saw **her** play baseball.	They saw **themselves** play baseball.	They saw **them** play baseball.
'None' response	He has no capacity to be a teacher.	none	They have no capacity to be a teacher.
Other modifications	In any case, I will tell him about the critical tone **your** House has adopted on this issue.	In any case, I will tell them about the critical tone **their** House has adopted on this issue.	In any case, I will tell them about the critical tone **your** House has adopted on this issue.

Table 9: Examples for the error labels described in Table 8

Output	Reference
They advised them to give up smoking, but they wouldn't listen.	They advised them to give up smoking, but they wouldn't listen to them.
They were able to hold back their anger and avoid a fight.	Jim was able to hold back their anger and avoid a fight.
The news that they had gotten injured was a shock to them.	The news that they had got injured was a shock to them.
They have done it with their colleagues and the Committee of Legal Affairs.	They have done it with the colleagues and the Committee of Legal Affairs.
In this respect , they have been very successful.	In this respect I believe that they have been very successful.
They cannot be older than me.	They cannot be older than I.
They suggested going to the theater, but there weren't any performances that night.	They suggested to go to the theater, but there weren't any performances that night.

Table 10: More examples of errors of type 'Other Modifications'. Differences are in red.

Category	Error Label	Zero-shot				Few-shot			
		tu	hu	fi	fa	tu	hu	fi	fa
Corrections	Comma	0	0	2	2	0	0	0	0
	Other Corrections	0	0	0	2	0	0	0	0
Neutral rewriting	POS	2	1	9	4	0	0	0	1
	SVA	5	5	3	0	0	9	0	2
	Them → Themselves	0	10	10	0	0	4	0	0
	'None' response	4	6	20	11	4	6	22	10
Irrelevant Modifications	Other modifications	16	16	33	8	4	10	32	10
Total		27	38	77	27	4	19	54	23

Table 11: Error analysis of GPT-3.5 Turbo's zero-shot and few-shot performance in English gender-neutral rewriting task.

Category	Feminine	Masculine	Neutral
Subject	She	He	They
Object	Her	Him	Them
Possessive Determiner	Her	His	Their
Possessive Pronoun	Hers	His	Theirs
Reflexive	Herself	Himself	Themselves

Table 12: Pronoun categories

he	she	him
her	his	himself
herself	Ms	Mrs
Ms.	Mrs.	madam
woman	women	actress
actresses	airwoman	airwomen
aunts	aunt	uncle
uncles	brother	brothers
boyfriend	boyfriends	girlfriend
girlfriends	girl	girls
bride	brides	sister
sisters	businesswoman	businesswomen
chairwoman	chairwomen	chick
chicks	mom	moms
mommy	mommies	grandmother
daughter	daughters	mother
mothers	female	females
gal	gals	lady
ladies	granddaughter	granddaughters
grandmother	grandmothers	grandson
grandsons	grandfather	grandfathers
wife	wives	queen
queens	policewoman	policewomen
princess	princesses	spokeswoman
spokeswomen	stepson	stepdaughter
stepfather	stepmother	stepgrandmother
stepgrandfather		

Table 13: Gendered English Nouns and Pronouns