

Gender Bias and Contextual Sensitivity in Machine Translation: A Focus on Turkish Subject-Dropped Sentences

Şeyda PORTILLO-PALMA* and Sergi ALVAREZ-VIDAL**

Turkish, a language that does not explicitly mark gender in pronouns, poses a unique challenge for machine translation systems, particularly in cases of gender-neutral or ambiguous context. This study investigates the performance of neural machine translation (NMT) and large language models (LLMs) in resolving gender ambiguity when translating Turkish subject-dropped sentences into English. The analysis examines four prominent models—Google Translate, DeepL, ChatGPT, and Gemini—evaluating their pronoun selection and the extent of gender bias, especially in emotionally charged or contextually nuanced sentences. A primarily quantitative evaluation reveals a persistent gender bias across all models, with LLMs demonstrating relatively better performance than NMTs when clearer contextual information is present. However, all models exhibit limitations in managing the complexities of cross-linguistic gender representation. This research highlights the pressing need for gender-neutral solutions and advancements in context-sensitive translation. Furthermore, we introduce a moderately sized annotated Turkish corpus, designed to facilitate future studies on gender ambiguity in machine translation (MT). This dataset provides a valuable resource for enhancing the accuracy of gendered pronoun resolution and fostering more inclusive, bias-reduced translation systems. Overall, the study contributes to the growing discourse on reducing bias in language models while addressing the challenges of nuanced linguistic diversity in translation.

Keywords: gender bias; emotion translation; machine translation; context awareness; anaphora resolution

1. Introduction

The Turkish language is spoken by over 80 million people, primarily in Türkiye, with significant speaker populations in Cyprus, Continental Europe, and Central Asia as well.¹ It features several unique linguistic characteristics that pose challenges for NLP applications, especially when compared to more extensively studied languages.

* Graduate student at Universitat Pompeu Fabra, Barcelona.

E-mail: seydanur.portillo01@estudiant.upf.edu; ORCID ID: <https://orcid.org/0000-0001-6986-231X>.

** Assistant professor at Universitat Autònoma de Barcelona.

E-mail: Sergi.Alvarez@uab.cat; ORCID ID: <https://orcid.org/0000-0002-6355-4559>.

(Received 19 October 2024; accepted 12 December 2024)

¹ This study was conducted based on the Turkish spoken in Türkiye.

Firstly, Turkish has neither grammatical gender as a paradigmatic feature of nouns, nor such gendered or inanimate personal pronouns as he/she/it in English. The only third person singular pronoun, *o*, may refer to ‘he,’ ‘she,’ or ‘it’ with its corresponding declined forms in table 1.

Table 1. Case declensions of 3SG *o* in Turkish

NOM	GEN	DAT	ACC	ABL	LOC
<i>o</i>	<i>onun</i>	<i>ona</i>	<i>onu</i>	<i>ondan</i>	<i>onda</i>

Consequently, determining the gender of a subject in a sentence without a specific context becomes challenging. This challenge is particularly pronounced when dealing with machine translation (MT) tasks. In such cases, gender bias can become apparent, especially when the target language lacks gender neutrality. Gender bias in MT models has been observed in various studies, highlighting persistent challenges in addressing these biases (Ciora, Iren, and Alikhani 2021). Sourojit Ghosh and Aylin Caliskan (2023) also found that ChatGPT’s translations between English and Bengali, as well as five other gender-neutral languages including Turkish, exhibited significant gender biases mainly in the translation of occupations (e.g., male doctor, female nurse) and actions (e.g., cooking woman, man going to work).

Yet, this is not the only manner in which bias can appear in MT. Turkish belongs to a typological category that allows the omission of subjects in sentence formation. This group of languages is known collectively as pro-drop languages, in which the omission of subject pronouns in any sentence is permitted without violating syntactic constraints. For instance, sentence (2) with a null subject conveys the same semantic information as (1) and is structurally sound, despite the omission of the explicit subject pronoun *ben* (I) which is present in (1).

- (1) *Ben ev-e gel-di-m.*
 I home-DAT come-PF.1SG²
 (I came home.)

² Glossary abbreviations in this study are adapted from Göksel and Kerslake (2005).

- (2) \emptyset *ev-e gel-di-m.*
home-DAT come-PF.1SG
(I came home.)

Likewise, in (3) all the relevant information about the grammatical person, number, negation, verb tense, and even evidentiality can be retrieved from the conjugated verb *gelmiyorlarmış*. This is because Turkish is agglutinating, meaning that it is a synthetic language that concatenates long strings of morphemes together without fusion. This, in turn, implies that Turkish has a large morphological paradigm and high word information density, also analyzable as low syllable information density (Bentz et al. 2023).


- (3) *Park-a gel-m-iyor-lar-mış.*
park-DAT come-NEG-IMPF.3PL-EV.PF
(Apparently they are not coming to the park.)

On the other hand, non-pro-drop languages cannot transmit verb-subject information within a single word. For instance, the translation of (3) to English employs ‘coming,’ transmitting only the value of the verb aspect as imperfective (IPFV). Therefore, to communicate the same grammatical values as in the Turkish *gelmiyorlarmış* within the constraints of a non-agglutinating language, there would have to be a bound form of the subject-pronoun (encoding both person and number) that affixes to the verb stem. To convey the totality of the information encoded in the Turkish verb, the verb stem would have to affix elements of negation and evidentiality as well. Since all the information the addressee needs to make sense of the expression is available in the verbal inflection suffixes, pro-dropping is used quite frequently in Turkish. This means that when translating from Turkish to English, the MT model must choose a pronoun—either *he*, *she*, or *they* as a gender-neutral alternative—to accurately convey the semantic information in the target language.

Figure 1. ChatGPT, attributing anger to men



Figure 2. ChatGPT, attributing sadness to women

 **You**
ağlıyordu

translate to english

 **ChatGPT**
"She was crying."

🔊 📄 ↺ 🗨️

Sentence (4b), as translated in figure 1, corresponds to the subject-dropped form (4a), which has the same translation into English as its non-drop counterpart.

- | | |
|---|---|
| <p>(4a) <i>O ağl-(i)yor-du</i>
3SG.SUBJ cry-IPFV-P.COP.3SG
(He/she was crying.)</p> | <p>(4b) \emptyset <i>ağl-(i)yor-du</i>.
cry-IPFV-
P.COP.3SG
(He/she was crying.)</p> |
|---|---|

Respectively, sentence (5b), as translated in figure 2 above, corresponds to the subject-dropped form (5a), which also has the same translation into English as its non-drop counterpart.

- | | |
|--|--|
| <p>(5a) <i>O çok sinirli-(y)di</i>.
He/she very angry-P.COP.3SG
(He/she was very angry.)</p> | <p>(5b) \emptyset <i>çok sinirli-(y)di</i>.
very angry-P.COP.3SG
(He/she was very angry.)</p> |
|--|--|

The omission of subject pronouns and the overall absence of overt gender-marking in the source language can pose challenges for MT tools, particularly when selecting pronouns in non-pro-drop target languages. In languages like Turkish, which lacks explicit gender pronouns and allows for the omission of subject pronouns (pro-drop), MT systems may struggle to accurately determine gendered pronouns in translation. In the example above, the language model attributes actions like ‘crying’ to women and ‘anger’ to men, reflecting potential gender stereotypes. These examples raise the question of whether coreferential bias—a tendency for translation systems to inaccurately resolve references to entities—is prevalent when translating from a pro-drop, gender-neutral language like Turkish into a non-pro-drop language with gender-specific pronouns.

This paper aims to explore whether such biases are persistent in MT tasks. We examine how the MT systems handle these cases and whether coreferential bias manifests in the translation process. As Turkish is classified among the low-resource languages and

characterized by a scarcity of annotated datasets for various task types in NLP (Aleçakır, Bölücü, and Can 2022), another essential objective of this study is to contribute to the field by providing a moderately sized annotated corpus in Turkish. Therefore, we have created a small dataset with sentences that exemplify scenarios where pronoun selection in the target language could be influenced by gender stereotypes. The dataset we created consists of Turkish sentences that express six different emotions—happiness, sadness, anger, fear, surprise, and disgust. Since Turkish does not have gendered pronouns, these sentences are gender-neutral in the source language. We tested four models—Google Translate, Gemini, DeepL, and ChatGPT—to analyze how they handle gender pronoun selection when translating these sentences into a gendered language.

Section 2 will provide an overview of MT and its challenges with respect to gender bias. In section 3, we detail the methodology we have followed for the dataset collection and MT evaluations, followed by the results in section 4 and the conclusions in section 5.

2. Machine Translation and Gender Bias

In recent years, MT has become very popular among users of all ages due to the improvement in the quality of NMT models. Furthermore, recent advances in LLMs, especially transformer-based models like BERT (Devlin et al. 2019), and Generative Pre-trained Transformer (GPT) have significantly enhanced the use of MT across industries and among individual users. These models have made translation services more accessible and affordable, driving their popularity even further.

Even though they are based on similar technologies, LLMs and NMT systems have several key differences (Vaswani et al. 2017). An encoder-decoder structure is employed in NMTs where the source sentence is encoded in the encoder network and the target sentence is decoded based on the previous outputs in the decoder network. They often rely on large amounts of parallel data (Hendy et al. 2023) and utilize contextual information from both the source and target languages to generate translations (Wang et al. 2022). This specialization often allows them to excel in capturing language-specific nuances and producing accurate translations. LLMs like GPT models, on the other hand, have decoder-only architectures that utilize the same parameters for processing both context and source as a single input for generating subsequent outputs. They are predominantly trained on monolingual data, with a significant emphasis on

English. These models require a substantially larger number of parameters to attain multilingual in-context capabilities (Hendy et al. 2023).

Despite their differences from NMTs in architecture and training data, LLMs have demonstrated promising translation capabilities. In a variety of language pairs, they can capture complex contextual dependencies and produce coherent text based on the given context (Castilho et al. 2023), though they are not specifically tailored for translation tasks. A case study conducted by Xinchun Li (2024) comparing LLMs and NMTs in Chinese-English translation and measuring the effect of genres and translation directions found that LLMs do not significantly differ from NMT systems in terms of translation quality. While commercial NMT systems excel at delivering accurate translations within specific domains or languages, LLMs are proficient in generating natural-sounding translations and handling rare words that NMT systems struggle to process (Zeng et al. 2024). Wang et al. (2023) conducted a comparative analysis of commercial MT systems alongside document-level NMT and GPT models, focusing on their discourse awareness. Their main finding is that ChatGPT performed better than the commercial MT systems. These discoveries imply a promising future for LLMs as proficient MT tools in the near future. Their contextual understanding can be beneficial in translation tasks, especially for handling ambiguous or context-dependent language constructs.

Alongside the advancements in translation technology, concerns about bias in MT have emerged. Addressing bias in MT has become a crucial area of research and development, aiming to create more inclusive and accurate translations that reflect the diversity of languages and cultures. Both LLMs and NMT tools can inadvertently perpetuate biases present in the data they are trained on, which may manifest in various forms, such as gender, race, or cultural biases, and can significantly impact the accuracy and inclusivity of translations. These biases can stem from various sources and they can be categorized into three types as proposed by Batya Friedman and Helen Nissenbaum (1996): (1) pre-existing biases, which stem from historical, social, and cultural contexts and are embedded in the data due to human behaviors, decisions, and societal norms; (2) technical biases introduced by the technical aspects of the design, development, and implementation of these systems; and (3) emergent biases that arise when a system is deployed and interacts with real-world environments and users in ways that were not anticipated during its development.

For NMTs, dealing with gender bias stands out as one of the toughest and most complex challenges. In an attempt to enhance user experience, Google Translate first has introduced a

feature to include both genders in its output when gender ambiguity is identified (Kuczmarski and Johnson 2018). This innovative approach made a significant leap forward in tackling gender bias and was later adapted by other service providers such as DeepL. In spite of various mitigation methods proposed for NMTs thus far, including the contributions of Basta (2022) and Costa-Jussà et al. (2022), a recent study by Piazzolla, Savoldi, and Bentivogli (2023) revealed the persisting issue of gender bias in the output of widely used NMT tools including DeepL and Google Translate. One of the possible solutions to this embedded issue is to rely on coreference resolution. This tactic has been elucidated by Sánchez et al. (2023), who find that coreference resolution to assign gender in translations ensures that LLMs exhibit broader gender variability in datasets with ambiguous contexts, while maintaining consistency in clearer ones. As of yet, this strategy is not a panacea. Eva Vanmassenhove (2024) reveals that even with explicit prompting, ChatGPT struggles with producing gender-inclusive translations in English when the source language is Italian, often omitting feminine and gender-neutral alternatives. The model's inability to handle gender systematically, as highlighted by missing gender-neutral markers in Italian, underscores a persistent bias. Similarly, Plaza-del-Arco et al. (2024) demonstrated how gender bias manifests in LLMs, leading to stereotypical emotional assignments by gender, such as 'angry man' and 'sad woman.' These examples highlight the importance of ongoing efforts to mitigate bias and improve the fairness of MT systems. Bias may be reflected or even amplified by machine learning tools, impacting the perception of the technology and reinforcing societal inequalities. As the complexity of handling gender bias in translation tasks largely stems from the disparity between social and linguistic gender categories (Stanczak and Augenstein 2021), it is crucial to integrate insights from disciplines beyond engineering.

Additionally, there is a significant lack of studies directly comparing LLMs with traditional NMT systems, especially in how they handle distinct linguistic patterns when translating between language pairs like English and Turkish. Admitting the fact that LLMs are making strides in advancement, it remains as a fact that they still confront a significant disparity when compared to commercial translation systems, particularly concerning languages with limited resources (Zhu et al. 2024). Translating from pro-drop to non-pro-drop languages is still a significant and challenging task, as highlighted by Wang et al. (2018). Although it has been studied and several mitigation methods have been proposed for pro-drop languages such as Italian, Spanish, Chinese, and Japanese (Russo, Loáiciga, and Gulati 2012; Wang et al. 2017;

Xu et al. 2022), it remains a largely unexplored area for the majority of the languages that share similar features.

In order to properly address these issues, it is imperative to adopt a multifaceted approach that not only adjusts algorithmic parameters and mechanisms but also confronts how different languages and cultural perceptions are represented in these systems.

3. Methodology

Despite including qualitative aspects in the interpretation of potential bias and manual annotation, the main body of the analysis relies largely on quantitative methods. For data processing and visualization, Python programming language has been used.

3.1 Data Creation and MT Engines

The tendency for emotions to be stereotypically attributed to gender becomes more meaningful when considered in conjunction with a vast body of studies claiming different patterns of emotional expression between men and women. Plant et al. (2000) observed that men are often depicted as more prone to emotions like anger and pride, while women are associated with feelings of awe, fear, guilt, happiness, and sadness. These findings reflect conventional gender stereotypes, portraying men as assertive and dominant, traits closely linked to pride and anger. Stephanie Shields (2013) further emphasizes this contrast by highlighting women's portrayal as nurturing and empathetic, qualities typically extrapolated to feelings of joy and sadness. In considering regional perspectives, insights from a study in Türkiye conducted by Nuray Sakallı-Uğurlu, Beril Türkoğlu, and Abdülkadir Kuzlak (2018) were also taken into account. Although not directly addressing gendered emotions, the study explored gender stereotypes in the Turkish context. It found that men were often described with such traits as strength, selfishness, ambition, anger, rudeness, rationality, bravery, protectiveness, inconsideration, and toughness. These traits, closely associated with emotions like anger and pride, reflect the perceived societal construction of masculinity in Türkiye. This integration of regional perspectives further elucidates the association between gender stereotypes and emotional expressions within different cultural contexts.

The emotions investigated in this study were chosen primarily based on the International Survey on Emotion Antecedents and Reactions (ISEAR; Scherer and Wallbott 1994) dataset

that originally classifies emotions as anger, disgust, fear, sadness, shame, joy, and guilt. ISEAR is previously used in several studies investigating emotion detection and bias in the AI and ML field (Odbal, Zhang, and Ananiadou 2022; Plaza-del-Arco et al. 2024; Asghar et al. 2020; Wegge and Klinger 2024). It is based on the six universal emotions proposed by Paul Ekman (1992), excluding ‘surprise’ from the original set and adding ‘shame’ and ‘guilt’ resulting in a set of seven: anger, disgust, fear, happiness, sadness, shame, and guilt. The categories of ‘disgust’ and ‘guilt’ are excluded for this study, considering ‘guilt’ as a subcategory of ‘shame’ due to limited verbal expressions around it in Turkish. The final list of emotions relevant to the data analyzed henceforth consists of six categories: anger, pride, fear, joy, sadness, and shame.

Using Sketch Engine, linguistic data were extracted from two parallel corpora in Turkish. The first reference corpus is OpenSubtitles 2011 which consists of 166,085,459 tokens. It is a sub-corpus of the Open Parallel Corpus (OPUS) – Turkish with 207,223,730 tokens and 31,148,523 sentences in total (Tiedemann 2012). The second corpus, OpenSubtitles 2018 parallel – Turkish contains 630,921,773 tokens and 114,126,315 sentences in total (Lison and Tiedemann 2016).

It is worth noting that there are certain differences between written and spoken Turkish. While written Turkish typically adheres to a formal structure, spoken Turkish tends to be more informal (Dursunoğlu 2006). Exemplifying the informality of the spoken register, subject-dropping is more prevalent in spoken Turkish than in the written form. Thus, movie subtitles are a preferred data source for the analysis of null-subject-sentences, as they reflect the informality and common patterns of spoken Turkish in a more nuanced and natural manner.

Sentences explicitly conveying emotions such as anger, pride, fear, sadness, shame, and joy were manually extracted from the aforementioned corpora. The newly created dataset contained 1,734 sentences. Although it was possible to expand the data set by increasing this number, it was predicted that no significant change in the general trend would be observed. Approximately 300 null-subject-containing sentences were set per emotion category. Using the parallel concordance feature of Sketch Engine (see figure 3), for each category of emotion, half of these 300 sentences were carefully chosen among the ones holding a covert feminine subject, while the other half had a covert masculine subject in the broader context which they were extracted from.

Figure 3. Parallel concordance of English and Turkish sentences expressing fear with the same pronoun

English Sentence	Turkish Sentence
OpenSubtitles2011 <s> If she 's one of them, she won't be as scared . </s>	Onlardan biri olursa korkmayacağını düş
OpenSubtitles2011 <s> It's like, uh, she 's scared of nothin' . </s>	Sanki hiçbir şeyden korkmuyor. </s>
OpenSubtitles2011 <s> She 's scared . </s>	Korkuyor. </s>
OpenSubtitles2011 <s> She looked pretty scared to me. </s>	Bence fazlasıyla korkmuştu. </s>
OpenSubtitles2011 <s> The roof got blown off and scared my wife so bad ... </s>	Çatım uçtu ayrıca karım da çok yak
OpenSubtitles2011 ... that she went into labor. </s>	doğuracak. </s>

In order to prevent overlapping, only one emotional reference per sentence was included. The sentences that include adjectives or occupations stereotypically attributed to certain genders were excluded, aiming to maintain the focus of the research solely on emotional attribution and analyze the bias associated with it.

The number of sentences collected per emotion has been counted (see figure 4) to ensure that the dataset is balanced and does not hold the tendency to overrepresent or underrepresent certain emotions. A similar procedure was followed to also have a balanced representation of gender pronouns (see figure 5).

Figure 4. Emotional representation

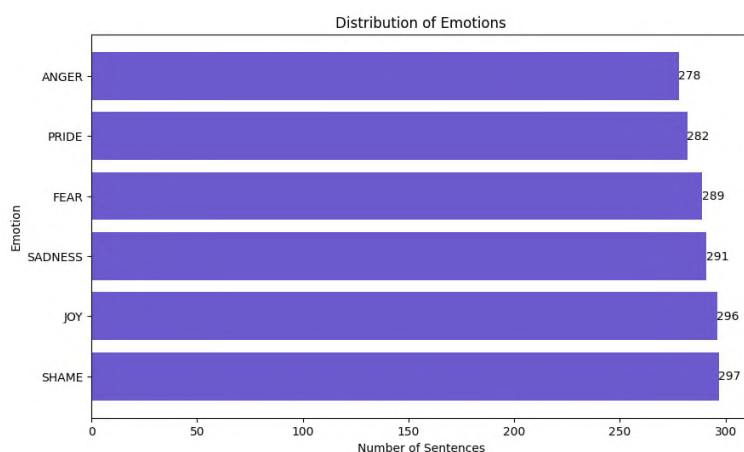
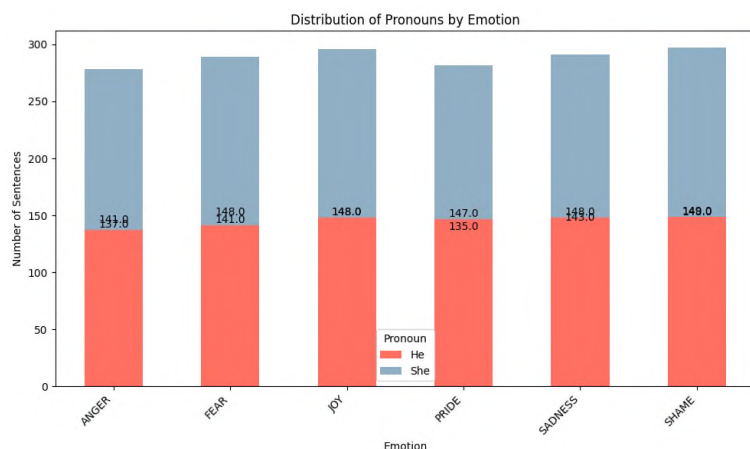


Figure 5. Gender distribution



Upon completing the balance check, the gender-specifying contextual clue that is tied to the null subject with anaphoric and cataphoric reference was added. Anaphoric reference is the semantic relationship that exists between any given word in a text or speech and any anterior element that the same word may refer to. Cataphoric reference happens when the direction of the anaphora is forward; in other words, when the word in a given context refers to another element that appears after itself. Figure 6 shows an instance of anaphora in which the pronoun (*she*) is prior to the referent (*her mother*), while figure 7 illustrates cataphora in which the pronoun (*she*) is posterior to the actual entity (*Ice Princess*).

Figure 6. Anaphoric relation between the ‘she’ pronoun and ‘her mother’

ous (OPUS) – English
t. </s></s> - I just may take you up on that. </s></s> - Lane' s fine. </s></s> - Good. </s></s> I haven
> - And? </s></s> Her mother answered. </s></s> She sounded angry . </s></s> - No, that' s just M
n time, she said she' d have the FBI trace the call ... and have me thrown in prison. </s></s> And th

Figure 7. Cataphoric relation between the ‘she’ pronoun and the ‘Ice Princess’

ply resentful. </s></s> Maybe you should retire. </s></s> Damn. </s></s> Outstanding! </s></s>
:.. </s></s> You said you' d scare the Ice Princess. </s></s> She looked pretty scared to me. </s></s>
</s></s> Let' s consummate our fiendish union. </s></s> What? </s></s> Oh, come on. </s></s>
> |'

In this way, a total of 1734 sentences were placed in a column without contextual information referring to pronoun gender, as is illustrated in figure 8.

Figure 8. Sentences without contextual information

⇒ DataFrame Structure:

	Sentence	Emotion	TRUE	PN_Gemini	\
0	Bu aralar o kadar huysuz ki, hatta bir iki ker...	ANGER	She	He	
1	Bu teşhisi koymamın nedeni esasında alkol deęi...	ANGER	She	He	
2	On kilo pirinç götürmek istedi. Haliyle, olmaz...	ANGER	She	He	
3	Onunla kızı hakkında konuştum ve çok sinirlendi .	ANGER	She	He	
4	Bekledięi için sinirlendi.	ANGER	She	He	

	PN_GPT	PN_Google	PN_DeepL	\
0	He	He	Binary	
1	He	He	Binary	
2	He	He	He	
3	He	He	He	
4	He	He	Binary	

The same sentences, then, were copied to a separate column in their broader context with anaphoric and cataphoric references that made the gender explicit. Later on, each sentence was manually annotated with its gold pronoun (he or she), the contextual information (CI), and the distance between the CI and the phrases that convey the emotion.

In some cases, terms such as *wife*, *husband*, *boyfriend*, and *girlfriend* were excluded as gender indicators, as they are not required to have a gendered possessive pronoun in Turkish.

(6) (TR) *Fusako* iyi mi acaba. </s><s> Kocasıyla barışmış olmalı. </s> <s> Geldiğinde biraz üzgündü.

(ENG) I wonder if **Fusako** is okay. She must have reconciled with (her) husband. </s><s> She was a little sad when she arrived.

In example (6), instead of *kocası* (his/her husband), *Fusako* was considered as the contextual clue, which is a Japanese female name and has a cataphoric reference to ‘she’ in the following sentence.

In the column where sentences are presented without CI, the sentence that contains the name *Fusako* was extracted to challenge the MT model to assign a subject pronoun in English, because the phrase *kocası* (her/his husband) does not have a gendered possessive pronoun in Turkish. Therefore, the model is expected to have half the chance of assigning masculine or feminine pronouns as having a husband is not unique to women.

Similarly, in (7) too, ‘kadınlar’ (women) is considered as the contextual clue, rather than ‘kocası’ (his/her husband).

(7) (TR) *Kocasını kaybetti, hakarete uğradı... Üzgün olması doğal. Sen karışma. Bu, kadınların arasında.*

(ENG) She lost (her) husband, was insulted... It’s natural for (her) to be sad. Don’t get involved. This is between **women**.

Additionally, gender-neutral (unisex) names were not considered as contextual indicators.

For reproducibility and use in future research, the complete annotated dataset has been publicly shared on Github.³

The whole dataset was translated with the two mentioned NMT models (Google Translate and DeepL), and the two LLMs (Gemini and ChatGPT). These four models are preferred for the study, based on their popularity in use and the ability to support translations between Turkish and English.

DeepL, launched in 2017 by DeepL SE, is a widely-used NMT that currently supports 32 languages. Just like most publicly available MT tools, DeepL also employs the Transformer architecture but with significant modifications. According to their website, DeepL distinguishes itself through a unique network architecture, advanced machine learning techniques that go beyond standard supervised learning and the efficient use of network parameters. Additionally, it focuses on targeted training data using specialized crawlers, unlike competitors that rely on vast amounts of general data (DeepL 2021).⁴

Overall, DeepL’s approach relies on learning patterns and structures from data rather than predefined linguistic rules.

Google Translate, released in 2006, supports 133 languages as of June 2024 and is one of the most widely used neural machine translation (NMT) systems. Initially launched as a statistical machine translation (SMT) model, in 2016 Google Translate announced its adoption of the neural-network-based architecture Google Neural Machine Translation (GNMT) (Wu et al. 2016; Turovsky 2016) which was then followed by the adoption of the multilingual NMT system facilitating ‘zero-shot’ translation.

³ https://github.com/seydaportillo/tr_emotbias_dataset.

⁴ <https://www.deepl.com/en/blog/how-does-deepl-work>.

ChatGPT is developed by OpenAI and first launched in 2022. It is a large-scale conversational model that's primarily designed for natural language understanding and generation rather than MT specifically. However, the model can be fine-tuned for a wide variety of tasks including translation and produce accurate output between the desired target language pairs based on user-provided prompts or the inputs. This enables the model to perform translation tasks by using the linguistic structures and patterns it has learned throughout training, despite not being specifically trained on it (Brown et al. 2020). With over 80 languages supported at the moment, ChatGPT is becoming more and more popular every day thanks to its regularly updated models.

Gemini AI, developed by Google DeepMind, is the second LLM included in this study. According to Google DeepMind, Gemini models can use long-context functions.⁵

Some versions can handle up to two million token context windows, making them very effective for processing long documents and complex queries. Based on MMLU benchmarks that they have been extensively tested on, Gemini models showed strong performance on a variety of subjects, though with specific biases in response selection sometimes (Ono and Morita 2024). At the time of writing, Gemini is available in 35 languages.

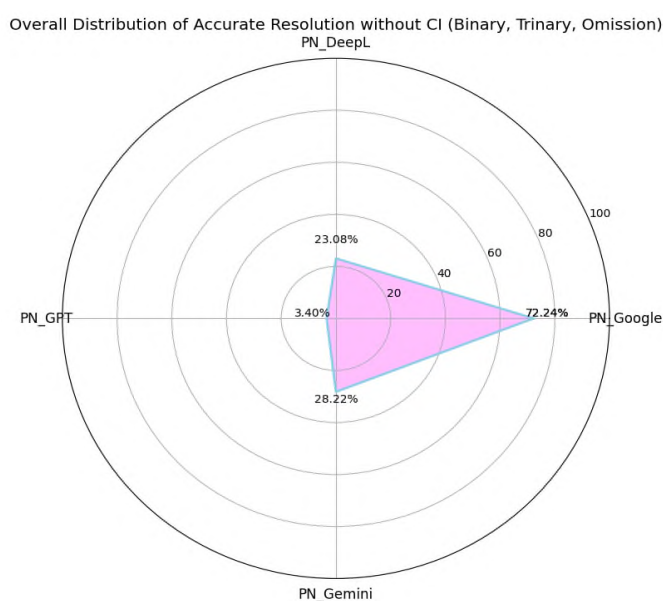
4. Results

For the sentences without context, the overall pronoun selection and their distribution for each model have been evaluated. The range of options across MT models included 'he,' 'she,' 'it,' 'they,' 'you,' 'we,' 'I,' 'NA,' 'omission,' 'binary' (he/she), and 'ternary' (he/she/it). Given that there is no context available to determine the gender of the subject, only the resolutions with 'omission,' 'binary' (he/she), and 'ternary' (he/she/it) have been mapped as correct resolutions. The rationale behind this mapping is that when the MT resolution is either 'binary' (he/she) or 'ternary' (he/she/it) the model is more likely to be sensitive to linguistic patterns where the subject pronoun lacks an explicit gender or is not present at all; therefore, it provides options to the user rather than arbitrarily assigning gender to the agent. An 'omission' indicates that the model is attempting to conform to the grammatical structure of the source language to the greatest extent possible given the differences between the target languages. All other model outputs have been classified as incorrect answers for the cases where further

⁵ <https://deepmind.google/technologies/gemini/>.

context is not present. For instance, using ‘he,’ ‘she,’ or ‘it’ without contextual clues is considered poor performance as the model lacks adequate information to accurately infer the subject’s gender. Translations such as ‘I,’ ‘you,’ ‘we,’ or the ‘NA’ demonstrate that the given model struggles in resolving a different kind of coreference problem, which is outside the scope of this study, as it focuses primarily on gender ambiguity.

Figure 9. Capability of handling null-subject sentences across models



Based on this metric, figure 9 indicates that with 72.24% accuracy, Google Translate has outperformed the rest of the models in terms of mitigating the challenges that gender-neutral pronouns or null-subject sentences pose in MT tasks. It is followed by Gemini and DeepL, with accuracy rates of 28.22% and 23.08%, respectively, indicating that the two models have substantially preferred single-gendered pronouns over Google Translate’s mitigation techniques. Chat GPT, on the other hand, showed the poorest performance among MT models. With 3.40%, the widely used LLM model employed the lowest number of mitigation techniques such as providing the user with optional translations (binary/ternary) or aligning the translation output with the input structure (omission).

After examining the overall translation results per MT tool, only ‘he’ and ‘she’ pronouns were counted for each model per emotion, and their percentages were calculated relative to one another. This method sought to uncover any patterns favoring certain gender pronouns for specific emotions, which could indicate a potential bias in the MT output.

As figure 10 shows, despite showing a great performance with 72.24% in mitigation, for the remainder, Google Translate favored masculine pronouns in all categories while translating null-subject sentences from Turkish to English.

Figure 10. Gendered pronoun selection of Google Translate per emotion category

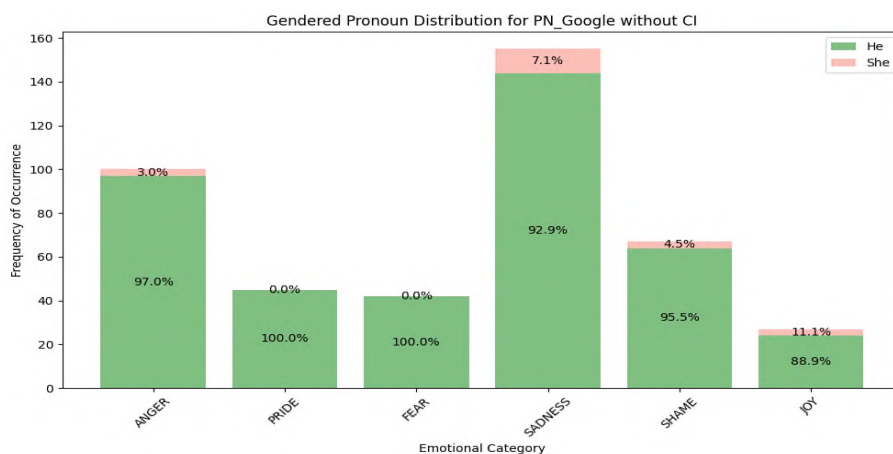
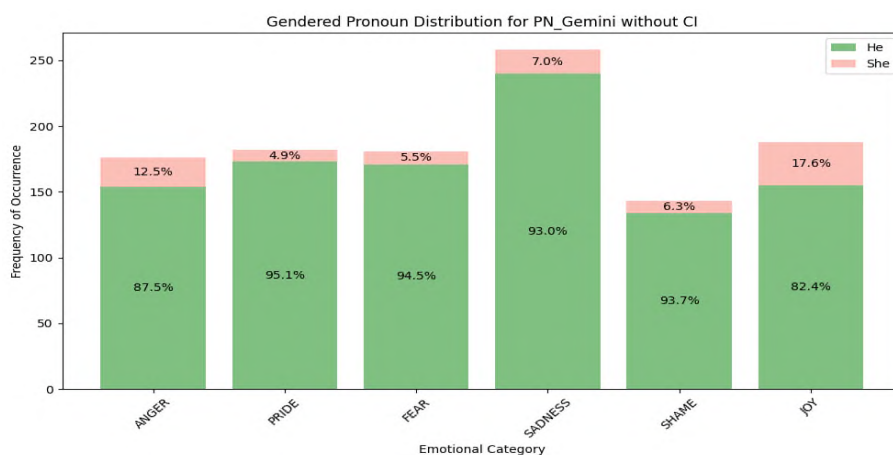


Figure 11 shows the distribution of ‘he’ and ‘she’ pronouns among the translations provided by Gemini. Similar to Google Translate, Gemini also shows a strong tendency to favor masculine pronouns in all categories.

Figure 11. Gendered pronoun selection of Gemini per emotion category



DeepL, on the other hand, shows a slightly different performance compared to Google Translate and Gemini. Although it largely follows the tendency to prefer the masculine pronoun over the feminine one, a significant decrease in masculine pronoun selection has been observed for ‘sadness.’ While ‘anger,’ ‘pride,’ and ‘fear’ retain more gendered assignments, DeepL is slightly more flexible in assigning feminine pronouns in sentences that convey such emotions

as ‘shame’ and ‘joy’ (see figure 12). These findings about DeepL’s pronoun selection in MT tasks align with the gender stereotypes and emotions that are often attributed to men and women in the society.

Similar to DeepL but even more precisely, figure 13 shows how ChatGPT also favored masculine pronouns nearly in 100% of the sentences that convey emotions such as ‘anger,’ ‘pride,’ and ‘fear’ that are often stereotypically attributed to masculinity in the society.

Figure 12. Gendered pronoun selection of DeepL per emotion category

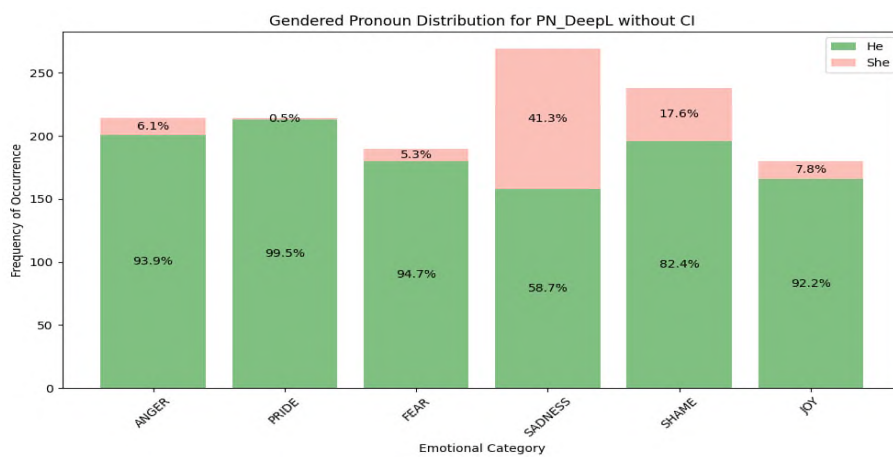
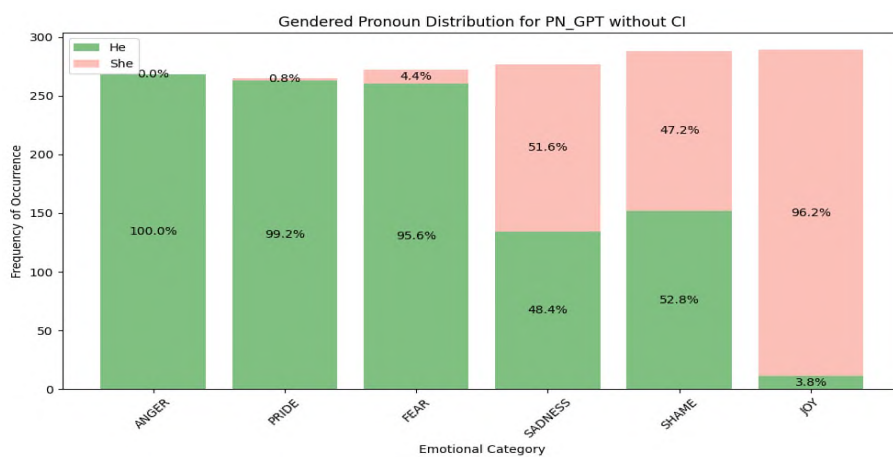


Figure 13. Gendered pronoun selection of ChatGPT per emotion category



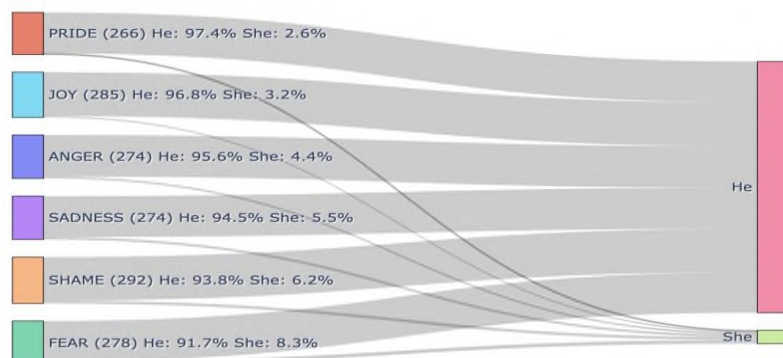
Overall, these findings indicate that all four MT models are biased to some extent and prefer masculine pronouns more often than feminine pronouns when further context is not available. In spite of the fact that some of them like Google Translate can effectively cope with the challenges caused by null-subject sentences or gender-neutral pronouns, in a broader picture

both Google Translate and DeepL favored masculine pronouns more often than feminine pronouns. This suggests that the two models demonstrate masculine bias although not specifically in emotion translation. ChatGPT and DeepL, on the other hand, are the only two models whose translation outputs align exactly with the gender stereotypes and societal expectations identified in various studies.

The next step aimed to evaluate the performance of the four models when broader context is available. Given that CI is available in the dataset as a clear indicator of gender, ‘binary’ (he/she) and ‘ternary’ (he/she) have not been included in the list of correct resolutions and only ‘he,’ ‘she,’ and ‘omission’ have been mapped as correct translations. Similarly to the first phase, only ‘he’ and ‘she’ pronouns for each translation tool per emotion have been counted. Their percentages relative to one another have been calculated, in order to be compared with the percentages of phase-one results and evaluated for a possible improvement. As stated during the data collection stages, about 150 sentences with covert female and 150 sentences with covert male pronouns were chosen for each emotion group. Thus, it was expected that if the subject pronoun was correctly detected based on CI, the representation of ‘she’ in the diagram would climb up and constitute half of the entire flow.

In accordance with this line of reasoning, when compared, figure 14 suggests a strong tendency for Google Translate to prefer masculine pronouns for all categories. Although initial findings suggest that Google Translate employs strategies to mitigate gender ambiguity, a closer examination reveals that it fails to process CI and understand language-specific nuances properly. This results in biased outputs when resolving semantic ambiguity in cross-linguistic tasks.

Figure 14. Gendered pronoun selection of Google Translate in clearer context



Therefore, Google Translate’s success in using a mitigation strategy when CI is not available indicates the model’s capability in dealing with null subjects rather than being an outright indicator of its overall translation performance dealing with bias.

Figure 15. Gendered pronoun selection of Gemini in clearer context

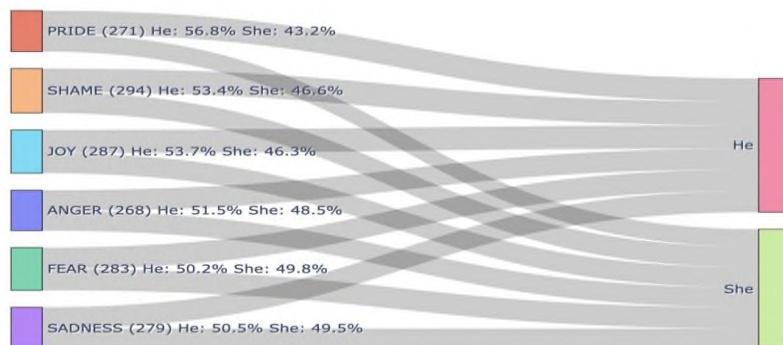
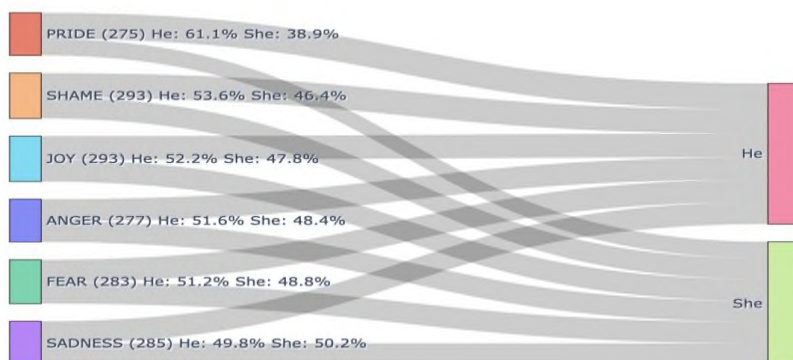


Figure 15 points out that Gemini boosted its performance in detecting the correct gendered pronoun for a given sentence, considering contextual details. The LLM model approaches nearly 50% translation accuracy in gender pronouns for all categories, suggesting that the model is able to process context correctly and somehow handles gender ambiguity in clearer context.

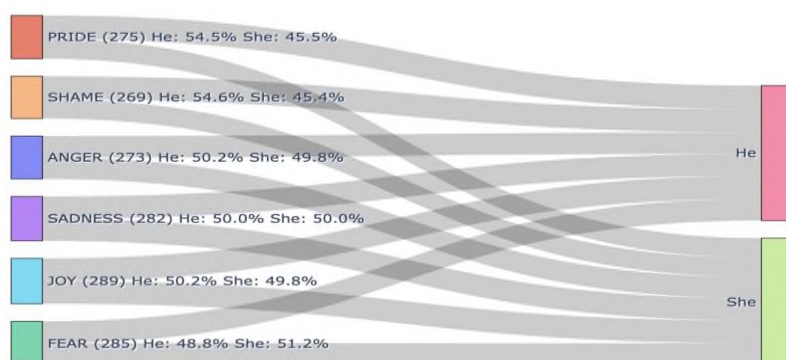
Another MT model whose performance in gendered-pronoun assignment improves with CI is DeepL. The difference between the percentage of feminine pronouns in the first and second phase results indicates that DeepL shows significant improvement at detecting the correct gender pronoun despite showing in the first phase dealing with null-subject sentences. Therefore, based on figure 16, it is plausible to say that DeepL output is much less biased when CI is given.

Figure 16. Gendered pronoun selection of DeepL in clearer context



Having shown the poorest performance with 3% of providing optional pronoun translations and showing masculine bias aligning with the gender stereotypes in the first phase evaluation without CI, ChatGPT appears to be the best performer in a clearer context. Approaching 50% in all categories, figure 17 shows that ChatGPT achieved great success in capturing contextual nuances.

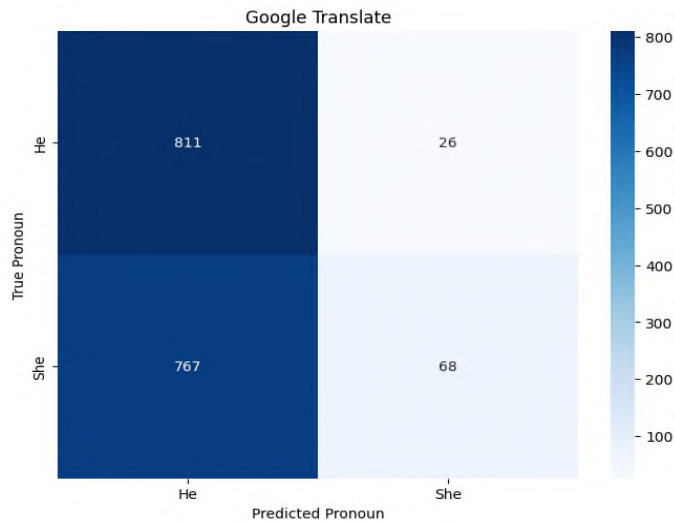
Figure 17. Gendered pronoun selection of ChatGPT in clearer context



At the final stage, MT pronouns have been compared to the annotated gold pronoun in the dataset. Four confusion matrices were generated for error analysis of each model and class-based F1 scores were computed to assess their overall ability to correctly detect ‘he’ and ‘she’ pronouns. In addition to ‘he’ and ‘she,’ ‘omission’ has also been mapped as the accurate resolution, regardless of the gold pronoun it corresponds with. This was ensured by creating a special function where ‘omission’ was counted equal to ‘he’ for cases in which the gold pronoun is ‘he’ and to ‘she’ for the ones with the gold pronoun ‘she.’

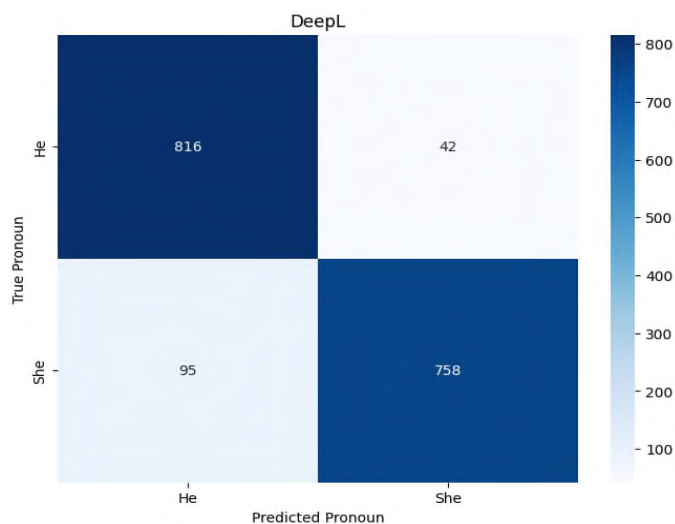
Based on the error analysis shown in figure 18, Google Translate showed 52.5% overall accuracy with F1 scores 0.669 for ‘he’ and 0.144 for ‘she’ pronouns.

Figure 18. Confusion matrix for Google Translate’s pronoun detection performance with CI



Similarly, figure 19 indicates that DeepL demonstrated 91.9% accuracy over the whole set with F1 scores 0.921 for ‘he’ and 0.915 for ‘she’ pronouns.

Figure 19. Confusion matrix for DeepL’s pronoun detection performance with CI



Gemini had an overall accuracy of 95.1%, with F1 scores of 0.951 for ‘he’ and 0.940 for ‘she’ pronouns based on the confusion matrix shown in figure 20.

Figure 20. Confusion matrix for Gemini’s pronoun detection performance with CI

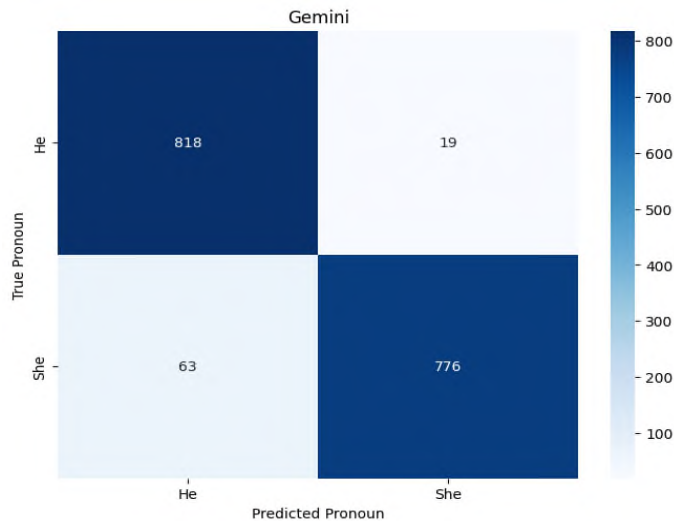
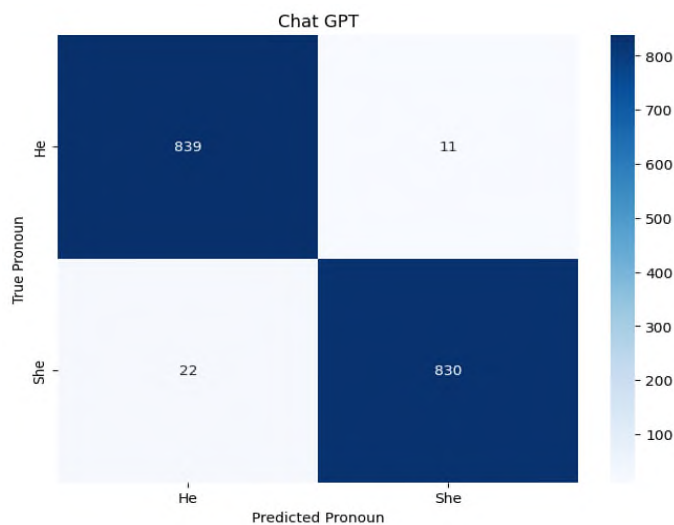


Figure 21. Confusion matrix for ChatGPT’s pronoun detection performance with CI



And finally, figure 21 presents the relevant data proving that ChatGPT acquired 98% overall accuracy with F1 Scores 0.980 for detecting ‘he’ and 0.993 for detecting ‘she’ pronouns, respectively.

In conclusion, Google Translate has much lower F1 scores for both pronouns compared to other models, particularly when recognizing ‘she’ pronouns. This could point to a bias or a specific deficiency in translating female pronouns. DeepL performs well with high F1 scores for both ‘he’ and ‘she’ pronouns, but not as well as LLMs Gemini and ChatGPT. Gemini does

really well, with slightly higher F1 scores than DeepL. Finally, ChatGPT gets the greatest F1 values for both pronouns, indicating the best overall performance of the four models.

5. Conclusion

The main motivation of this research was assessing the performance of NMTs and LLMs in resolving gender ambiguity and comparing their potential in perpetuating bias in MT. Based on the findings, LLMs like Gemini and ChatGPT outperform NMTs like Google Translate and DeepL in their context awareness, which is consistent with the findings of Wang et al. (2023). The findings are also consistent with Sánchez et al. (2023) regarding LLMs' ability to maintain gender consistency in clear contexts, though the broader gender variability in ambiguous contexts that they reported has not been observed in the case of Turkish. This is because, when broader context is not available, translations of DeepL (an NMT) and ChatGPT (an LLM) attributed emotions 'anger' and 'pride' to men and 'sadness,' 'shame' and 'joy' to women, which aligns with the perceived gender stereotypes towards women and cultural constructions of masculinity in Türkiye demonstrated in previous research (Shields 2013; Plant et al. 2000; Sakallı-Uğurlu, Türkoğlu, and Kuzlak 2018). Although Gemini does not follow this exact pattern as another LLM, the output produced by the model carries an overall masculine bias with its tendency to prefer masculine pronouns much more frequently than feminine pronouns. These findings, as a whole, underscore the ongoing concern about gender bias in LLMs, particularly in stereotypical emotional assignments. On the other hand, Google Translate, an NMT, is the only model that provides the user with alternative pronouns. The rest of the models, including LLMs, have poorer performance in mitigating gender ambiguity when contextual information is not provided, though they perform better in clearer context.

Google Translate's mitigation method must be adopted by other MT models and LLMs in resolving the challenges caused by null-subject sentences and gender-neutral pronouns. However, providing gender options alone is not effective without real improvement both in contextual understanding and cross-linguistic awareness across all types of models, whether NMTs or LLMs. Increasing the sensitivity of machine learning models in these aspects is crucial for increasing translation quality and decreasing bias amplification, hence boosting global communication.

References

- Aleçakır, Hüseyin, Necva Bölücü, and Burcu Can. 2022. “TurkishDelightNLP: A Neural Turkish NLP Toolkit.” In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, 17–26. Association for Computational Linguistics. doi:10.18653/v1/2022.naacl-demo.3.
- Asghar, Muhammad Zubair, Fazli Subhan, Muhammad Imran, Fazal Masud Kundi, Adil Khan, Shahboddin Shamsirband, Amir Mosavi, Peter Csiba, and Annamaria R. Varkonyi Koczy. 2020. “Performance Evaluation of Supervised Machine Learning Techniques for Efficient Detection of Emotions from Online Content.” *Computers, Materials & Continua* 63 (3): 1093–1118. doi:10.32604/cmc.2020.07709.
- Basta, Christine Raouf Saad. 2022. “Gender Bias in Natural Language Processing.” PhD diss., Universitat Politècnica de Catalunya. doi:10.5821/dissertation-2117-379361.
- Bentz, Christian, Ximena Gutierrez-Vasques, Olga Sozinova, and Tanja Samardžić. 2023. “Complexity Trade-Offs and Equi-Complexity in Natural Languages: A Meta-Analysis.” In “Measuring Language Complexity,” edited by Katharina Ehret, Aleksandrs Berdicevskis, Christian Bentz, and Alice Blumenthal-Dramé. Special Issue, *Linguistics Vanguard* 9 (1): 9–25. doi:10.1515/lingvan-2021-0054.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, edited by Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, and Maria-Florina Balcan. Curran Associates. doi:10.48550/arXiv.2005.14165.
- Castilho, Sheila, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. “Do Online Machine Translation Systems Care for Context? What About a GPT Model?” In *Proceedings of 24th Annual Conference of the European Association for Machine Translation*, edited by Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, et al., 393–417. European Association for Machine Translation. <https://aclanthology.org/2023.eamt-1.39>.
- Ciora, Chloe, Nur Iren, and Malihe Alikhani. 2021. “Examining Covert Gender Bias: A Case Study in Turkish and English Machine Translation Models.” In *Proceedings of 14th International Conference on Natural Language Generation*, edited by Anya Belz, Angela Fan, Ehud Reiter, and Yaji Sripada, 55–63. Association for Computational Linguistics. doi:10.18653/v1/2021.inlg-1.7.
- Costa-jussà, Marta R., Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2022. “Interpreting Gender Bias in Neural Machine Translation: Multilingual Architecture Matters.” In *Proceedings of the AAAI Conference on*

- Artificial Intelligence* 36 (11): 11855–11863. Association for the Advancement of Artificial Intelligence. California: AAAI Press. doi:10.1609/aaai.v36i11.21442.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by Jill Burstein, Christy Doran, and Tamar Solorio, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1423.
- Dursunoğlu, Halit. 2006. “Türkiye Türkçesinde Konuşma Dili ile Yazı Dili Arasındaki İlişki.” [The relationship between spoken language and written language in Turkish from Türkiye.] *Atatürk Üniversitesi Türkiyat Araştırmaları Enstitüsü Dergisi* 12 (30): 1–21. <https://dergipark.org.tr/tr/pub/ataunitaed/issue/2869/39214>.
- Ekman, Paul. 1992. “An Argument for Basic Emotions.” *Cognition and Emotion* 6 (3-4): 169–200. doi:10.1080/02699939208411068.
- Friedman, Batya, and Helen Nissenbaum. 1996. “Bias in Computer Systems.” *ACM Transactions on Information Systems (TOIS)* 14 (3): 330–347. doi:10.1145/230538.230561.
- Ghosh, Sourojit, and Aylin Caliskan. 2023. “ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages.” In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES 23)*, edited by Francesca Rossi, Sanmay Das, Jenny Davis, Kay Firth-Butterfield, and Alex John, 901–912. New York: Association for Computing Machinery. doi:10.48550/arXiv.2305.10510.
- Göksel, Aslı, and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. 1st ed. London: Routledge.
- Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young J. Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. “How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.” arXiv:abs/2302.09210. doi:10.48550/arXiv.2302.09210.
- Kuczumarski, James, and Melvin Johnson. 2018. “Gender-Aware Natural Language Translation.” *Technical Disclosure Commons*. https://www.tdcommons.org/dpubs_series/1577.
- Lison, Pierre, and Jörg Tiedemann. 2016. “OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles.” In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC’16)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios

- Piperidis, 923–929. European Language Resources Association (ELRA). <https://aclanthology.org/L16-1147>.
- Li, Xinchun. 2024. “Comparison of Translation Quality between Large Language Models and Neural Machine Translation Systems: A Case Study of Chinese-English Language Pair.” *International Journal of Education and Humanities* 4 (2): 121–128. doi:10.58557/(ijeh).v4i2.213.
- Odbal, Guanhong Zhang, and Sophia Ananiadou. 2022. “Examining and Mitigating Gender Bias in Text Emotion Detection Task.” *Neurocomputing*, no. 493, 422–434. doi:10.1016/j.neucom.2022.04.057.
- Ono, Kensuke, and Akira Morita. 2024. “Evaluating Large Language Models: ChatGPT-4, Mistral 8x7B, and Google Gemini Benchmarked Against MMLU.” *TechRxiv*. doi:10.36227/techrxiv.170956672.21573677/v1.
- Piazzolla, S. Alma, Beatrice Savoldi, and Luisa Bentivogli. 2023. “Good, but not always Fair: An Evaluation of Gender Bias for three Commercial Machine Translation Systems.” *HERMES - Journal of Language and Communication in Business*, no. 63, 209–225. doi:10.7146/hjlc.vi63.137553.
- Plant, E. Ashby, Janet Shibley Hyde, Dacher Keltner, and Patricia G. Devine. 2000. “The Gender Stereotyping of Emotions.” *Psychology of Women Quarterly* 24 (1): 81–92. doi:10.1111/j.1471-6402.2000.tb01024.x.
- Plaza-del-Arco, Flor Miriam, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. “Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution.” arXiv preprint. doi:10.48550/arXiv.2403.03121.
- Russo, Lorenza, Sharid Loáiciga, and Asheesh Gulati. 2012. “Italian and Spanish Null Subjects. A Case Study Evaluation in an MT Perspective.” In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 1779–1784. http://www.lrec-conf.org/proceedings/lrec2012/pdf/813_Paper.pdf.
- Sakallı-Uğurlu, Nuray, Beril Türkoğlu, and Abdülkadir Kuzlak. 2018. “How are Women and Men Perceived? Structure of Gender Stereotypes in Contemporary Turkey.” *Nesne Psikoloji Dergisi* 6 (13): 309–336. doi:10.7816/nesne-06-13-04.
- Sánchez, Eduardo, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2023. “Gender-specific Machine Translation with Large Language Models.” arXiv preprint. doi:10.48550/arXiv.2309.03175.
- Scherer, Klaus R., and Harald G. Wallbott. 1994. “Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning.” *Journal of Personality and Social Psychology* 66 (2): 310–28. doi:10.1037/0022-3514.66.2.310.

- Shields, Stephanie A. 2013. “Gender: An Intersectionality Perspective.” *Sex Roles* 68 (11–12): 675–689. doi:10.1007/s11199-008-9501-8.
- Stanczak, Karolina, and Isabelle Augenstein. 2021. “A Survey on Gender Bias in Natural Language Processing.” arXiv preprint. doi:10.48550/arXiv.2112.14168.
- Tiedemann, Jörg. 2012. “Parallel Data, Tools and Interfaces in OPUS.” In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC’2012)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2214–2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.
- Turovsky, Barak. 2016. “Found in Translation: More Accurate, Fluent Sentences in Google Translate.” *Google Blog*. November 15. <https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/>.
- Vanmassenhove, Eva. 2024. “Gender Bias in Machine Translation and The Era of Large Language Models.” arXiv preprint. doi:10.48550/arXiv.2401.10016.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention is All You Need.” In *Advances in Neural Information Processing Systems 30. 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*, edited by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, Vishy Vishwanathan, and Roman Garnett, 5999–6009. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
- Wang, Haifeng, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. “Progress in Machine Translation.” *Engineering* 18 (11): 143–153. doi:10.1016/j.eng.2021.03.023.
- Wang, Longyue, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. “Document-level Machine Translation with Large Language Models.” In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, edited by Houda Bouamor, Juan Pino, and Kalika Bali, 16646–16661. Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.1036.
- Wang, Longyue, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. “Translating Pro-Drop Languages with Reconstruction Models.” *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1): 4937–4945. doi:10.1609/aaai.v32i1.11913.
- Wang, Longyue, Zhaopeng Tu, Xiaojun Zhang, Siyou Liu, Hang Li, Andy Way, and Qun Liu. 2017. “A Novel and Robust Approach for Pro-Drop Language Translation.” *Machine Translation* 31 (1–2): 65–87. doi:10.1007/s10590-016-9184-9.

- Wegge, Maximilian, and Roman Klinger. 2024. “Topic Bias in Emotion Classification.” In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, edited by Rob van der Goot, JinYeong Bak, Max Müller-Eberstein, Wei Xu, Alan Ritter, and Tim Baldwin, 89–103. <https://aclanthology.org/2024.wnut-1.9/>.
- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. 2016. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.” *Computing Research Repository (CoRR)*. doi:10.48550/arXiv.1609.08144.
- Xu, Mingzhou, Longyue Wang, Derek F. Wong, Hongye Liu, Linfeng Song, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2022. “GuoFeng: A Benchmark for Zero Pronoun Recovery and Translation.” In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, edited by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, 11266–11278. Association for Computational Linguistics. doi:10.18653/v1/2022.emnlp-main.774.
- Zeng, Jiali, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. “Improving Machine Translation with Large Language Models: A Preliminary Study with Cooperative Decoding.” arXiv preprint. doi:10.48550/arXiv.2311.02851.
- Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. “Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis.” In *Findings of the Association for Computational Linguistics (NAACL 2024)*, edited by Kevin Duh, Helena Gomez, and Steven Bethard, 2765–2781. Association for Computational Linguistics. <https://aclanthology.org/2024.findings-naacl.176/>.