JoLIE 18:3/2025
doi: https://doi.org/10.29302/jolie.2025.18.3.6

GENDER RESOLUTION IN GOOGLE TRANSLATE AND CHATGPT: A DESCRIPTIVE CROSS-LINGUISTIC ANALYSIS

Ana-Maria Oprea
1 Decembrie 1918 University of Alba Iulia, Romania

Abstract

This study presents a descriptive cross-linguistic analysis of gender resolution patterns in two widely used machine translation systems, Google Translate and ChatGPT. Using a controlled challenge-set methodology, the analysis examines how each system assigns grammatical gender when translating gender-ambiguous source sentences into five typologically diverse target languages.
The test set comprises 175 constructed sentences designed to probe linguistic environments where gender resolution is known to vary, including occupational nouns, pronoun ambiguity, adjective-based descriptions, grammatical agreement, epicene nouns, prestige-related terms, and translations from a gender-neutral source language. Translation outputs were manually coded and analysed for distributional patterns across systems and languages.
The results document a recurrent tendency toward masculine default forms in gender-ambiguous contexts across both systems, with variation depending on target language and linguistic parameter. While ChatGPT more frequently provides alternative gendered renderings, its primary outputs show distributional patterns comparable to those observed in Google Translate. Cross-linguistic comparison suggests that typological features appear to influence how gender is resolved but do not eliminate default patterns.
This study is descriptive in scope and reports observed translation outputs under specific testing conditions. It does not aim to establish causal explanations or statistically generalisable claims about underlying system mechanisms beyond the tested sentence sets. Findings are limited to the tested sentence sets, systems, and time of evaluation and are not intended to support causal or generalisable claims about system design.

Key words: Gender bias; Machine translation; Translation studies; Cross-linguistic analysis; Controlled evaluation.

Received: 13 September 2025
Revised: 28 October 2025
Accepted: 12 November 2025
Published: 15 December 2025

1 Introduction

Gender "is perhaps the only grammatical category that ever evoked passion-and not only among linguists" (Matasović, 2004, p. 13, as cited in Corbett, 2014, p. 87). Because languages encode gender through distinct grammatical systems, lexical

conventions, and cultural histories, translation becomes a site for examining how gender is constructed, negotiated, and reproduced. When translation is delegated to machine learning systems, these linguistic complexities intersect with statistical inference, raising critical questions about how automated systems assign gender in contexts where human translators would exercise interpretive judgment.

Concerns about algorithmic gender bias have intensified in recent years, following high-profile cases in which AI systems reproduced or amplified historical inequalities. Amazon's automated recruiting tool was abandoned after it systematically penalised resumes that included the word *women* and downgraded graduates of all-women colleges. (Dastin, 2018). In 2019, CNN exposed a gender bias controversy involving Apple Card's credit limit assignments. (Nedlund, 2019) Users, including Apple co-founder Steve Wozniak, reported that women received lower credit limits than men, even when they shared financial profiles with male partners. The issue raised concerns about algorithmic bias in the assessment model, managed by Goldman Sachs. Despite denials of intentional discrimination, the incident sparked investigations into the role of AI algorithms in perpetuating gender bias. An experiment carried out in 2020 showed how Google Translate systematically changed the gender of translation and often defaulted to masculine terms in European languages, reinforcing stereotypes and erasing feminine forms in certain professions (like *historian* or *nurse*). This bias partly resulted from "language bridging" through English, which lacks gendered nouns, and it skews translations in languages with gendered structures. (Kayser-Brill, 2020)

These incidents point out a broader structural problem: MT systems reproduce linguistic patterns present in their training data, which may result in systematic gendered representations in output. In machine translation, where grammatical gender, semantic gender, and cultural associations intersect, the representational consequences of default gender assignment are a recurring concern in prior work. Prior research has demonstrated that MT systems default to masculine forms in ambiguous contexts, misgender feminine referents, and reproduce stereotypical associations between gender and occupation. Yet much of this work relies on corpus-based evaluation or focuses on a single language pair, leaving open questions about cross-linguistic consistency and the influence of architectural design.

The present study adopts a descriptive, comparative approach to examining how two widely used machine translation systems, Google Translate and ChatGPT, resolve grammatical gender in contexts where the source language provides limited or no explicit gender cues. By testing a controlled set of constructed sentences across multiple linguistic parameters and target languages, the study documents distributional patterns in gender assignment and compares how these patterns vary across systems and language types. Rather than proposing explanatory or causal accounts of system behaviour, the analysis focuses on observed outputs under specified testing conditions, with the aim of providing a systematic empirical record that can inform further methodological and theoretical work.

2 Literature Review

The question of gender bias in machine translation touches upon three research traditions: computational linguistics, translation studies, and gender studies. This section synthesises key findings from each, identifying the specific theoretical and methodological gaps this study addresses.

Languages encode gender differently, creating what linguistics calls typological variation in gender-marking systems. Some languages classify all nouns grammatically, assigning masculine, feminine, or neuter categories even to inanimate objects. In Hindi, for example, *the sun* is masculine while *the river* is feminine; in German, *the sun* is feminine and *the river* is masculine. Linguists believe these classifications evolved arbitrarily over time and are unique to each language. (Comrie, 1999, Kramer, 2014, Ghosh & Caliskan, 2023). Other languages, such as English and Swedish, have significantly less markers of grammatical gender except in pronominal systems, while Turkish and Finnish maintain minimal gender distinctions across their morphosyntactic structures (Corbett, 2014).

Beyond grammatical gender, languages also exhibit semantic or natural gender, where different lexical items distinguish male and female referents. English demonstrates this in pairs like *bull/cow*, *actor/actress*, and *lion/lioness*. Jakobson's (1972) concept of markedness illustrates a critical asymmetry in such pairs: the masculine form typically serves as the unmarked, default term, while the feminine is marked through morphological derivation or lexical specification. This linguistic pattern reflects and reinforces broader social hierarchies, as Gygax et al. argue: language acts "not only as a vehicle for beliefs, but also as a tool that builds them." (2019, p. 1).

These differences can create translation challenges: rendering a gender-neutral Turkish sentence into gendered Spanish requires the (human or machine) translator to make choices the source text leaves unspecified. Critically, these are not merely technical choices. Butler's (1999, p. 33) foundational concept of gender as "performative" established that linguistic forms do not simply reflect pre-existing gender categories but actively constitute them through repeated citation and enactment. Applied to translation, this means that choosing a term does not only mean adding a label to a referent, but participating in constructing gendered professional identities and reinforcing or challenging associations between gender and occupation. Baker (2006) extends this argument specifically to translation practice, positioning translators as active agents in narrative circulation rather than neutral conduits. As she argues, "translators and interpreters are responsible for the narratives they help circulate, and for the real-life consequences of giving these narratives currency and legitimacy" (Baker, 2006, p. 139). When MT systems default to masculine forms for high-status professions, they do not just make errors; they participate in reproducing status hierarchies and shaping public perception of who legitimately occupies particular social roles.

These perspectives motivate the topic and inform the discussion, but the present study operationalises gender resolution using descriptive coding of MT

outputs rather than theory-driven measures. The question is not simply whether MT systems produce grammatically correct output, but which realities they make visible.

From a historical perspective, the recognition that language choices carry political weight has deep precedent. Three examples from the drafting of the Universal Declaration of Human Rights illustrate how translation and linguistic formulation directly shaped international human rights discourse. Hansa Mehta, India's delegate to the UN Commission on Human Rights (1947-1948), successfully advocated for replacing "All men are born free and equal" with "All human beings are born free and equal" in Article 1, arguing that masculine generics obscured women's inclusion (United Nations 2019). Similarly, Minerva Bernardino, an advocate for human rights and gender equality, ensured that the preamble referenced "equal rights of men and women" rather than "equal rights of men", while Belorussian Rapporteur of the Commission on Human Rights in 1947, Evdokia Uralova contributed to Article 32's gender-neutral formulation: "Everyone, without any discrimination, has the right to equal pay for equal work" (Robert & Ethel Kennedy Human Rights Center 2023). These interventions demonstrate that linguistic choices are never merely stylistic but carry profound social and political weight. When masculine forms are privileged as default, they reinforce structural inequalities. In translation, this tendency becomes even more pronounced because translators must actively decide whether to reproduce, neutralise, or resist biases embedded in source or target languages. Gender bias in translation can thus be understood as "favouritism towards or prejudice against a particular gender" (bab.la 2024) that manifests when linguistic forms or translation choices perpetuate stereotypes, restrict representation, or erase identities.

The development of neural machine translation has not eliminated gender bias; in many cases, it has systematised and scaled it. Empirical research has established that contemporary MT systems systematically default to stereotypical gender assignments when source-language gender is ambiguous or neutral.

One of the most widely documented forms of gender bias in MT involves occupation-related terms. Prates et al. (2020) conducted a multilingual study translating sentences containing job titles from 12 gender-neutral languages into English using Google Translate. They found that the system exhibited "a strong tendency towards male defaults" (p.1) with translations overwhelmingly assigning masculine pronouns to high-status occupations (doctor, engineer, scientist, CEO, teacher) and feminine pronouns to lower-status or care-oriented roles (nurse, baker, wedding organiser, cleaner). Stanovsky et al. (2019) developed an automatic evaluation method using challenge sets for eight grammatical-gender languages (Spanish, French, Italian, Russian, Ukrainian, Hebrew, Arabic, German). Their dataset of 3.888 examples was balanced between male and female genders and stereotypical versus non-stereotypical gender-role assignments. Results indicated that all tested systems exhibited gender bias, with performance degrading particularly for feminine referents in counter-stereotypical contexts (e.g., female engineers, male nurses). The systematic nature of these patterns suggests they are not random errors but reflect embedded statistical associations in the training data.

While occupation-related bias has received significant attention, Savoldi et al. (2022) demonstrated that gender bias affects broader morphosyntactic structures. Their study introduced two new annotation layers to the MuST-SHE corpus: part-of-speech tags (POS) and agreement chains (i.e., sequences of words that must agree in gender within a sentence). This allowed them to test gender bias at a more granular level across speech translation systems. Their findings revealed that certain parts of speech-particularly adjectives and participles-are especially prone to errors in gender agreement. For example, in English-Italian translation, the sentence *The nurse said she was ready* was frequently rendered as *L'infermiere ha detto che era pronto* (using masculine forms for both noun and adjective) despite the feminine pronoun *she* in the source. The study also compared segmentation strategies, finding that character-based neural translation models handled gender agreement more consistently than byte-pair encoding (BPE) models, as they better preserved morphological differences in gendered word forms like *pronto* (masculine *ready*) versus *pronta* (feminine *ready*) in Italian. Savoldi et al. conclude that gender bias is „a problematic phenomenon affecting language technologies, with recent studies underscoring that it might surface differently across languages" (2022, p. 1). By highlighting variation across morphosyntactic structures rather than focusing solely on occupational nouns, their study emphasises the need for evaluation methods that reflect the entire complexity of gender expression in language. Gendered language patterns and translation discrepancies reflect intersections between linguistic structure, cultural norms and societal expectations.

A critical finding emerging from recent research is that gender bias manifests differently depending on language typology and the directionality of translation. Ghosh and Caliskan (2023) examined ChatGPT's translation capabilities across Bengali and five other low-resource languages, finding that the system perpetuated gender bias and ignored non-gendered pronouns in the GPT iteration. Their work highlighted that bias is not confined to high-resource European languages but extends to diverse linguistic contexts and concluded that these biases stem from historical and societal stereotypes embedded in training data. Vanmassenhove et al. (2018) demonstrated that adding explicit gender features to neural MT systems can materially reduce errors, noting that „adding a gender feature to an NMT system significantly improves the translation quality" on gendered phenomena across some language pairs (2018, p. 3003). This suggests that bias is partly addressable through architectural modifications, though such solutions require design choices and metadata not universally available across all language pairs or deployment contexts.

The directionality of bias also varies. Kayser-Brill (2020) documented how Google Translate systematically changed gender in translations involving European languages, often defaulting to masculine terms and erasing feminine forms in professions like *historian* or *nurse*. This bias resulted partly from a bridging process through English, which distorted gender translations in other.

Savoldi et al. (2021) identified the need for a unified framework to ease future research in this field. For this purpose, they relied on Friedman and Nissebaum's categorisation of machine bias sources (Savoldi et al. 2021, pp. 849-

50) and identified three types: (a) Pre-existing bias: rooted in historical, socio-cultural contexts and existing disparities in data. For example, if training corpora contain more instances of *male doctor* than *female doctor* due to historical professional demographics, systems learn these statistical patterns. (b) Technical bias: derived from technical constraints and decisions related to data creation, model design, and training/testing procedures. Vanmassenhove et al.'s (2018) post-editing studies showed that feminine referents require disproportionately more manual correction than masculine ones in MT workflows, increasing both cognitive and economic costs. (c) Emergent bias: occurring from the interaction between systems and users, especially when systems are used in contexts different from design environments.

The following "parameters" were addressed in the assessment of gender bias:

1. Representational harms made up of under-representation and stereotyping. Under-representation includes the reduction of visibility of certain social groups through language: misrepresentation of feminine entities as male in translation (default to masculine), lack of recognition for certain groups (e.g.: feminine entities translated as male, no account for gender neutral forms) or failure to reflect the identity and communicative repertoires. One example of the latter, in German-English translation refers to women who introduce themselves with the professional title *Ärztin* (*female doctor*) often translated simply as *doctor*, which erases the explicit female marking of their professional identity. While doctor in English is technically gender-neutral, the historical default association with maleness often results in readers assuming a male referent. (Hellinger, & Bussmann 2001).

Stereotyping involves the propagation of negative generalisations of a social group, such as: associating feminine representation with certain occupations (e.g.: teacher – feminine, professor – masculine) or associating feminine representation with attractiveness judgments. Collocation studies on the British National Corpus showed how women were more likely to be assigned adjectives that signal physical attractiveness, such as *pretty*, *attractive*, *beautiful*, and *pleasant-looking*. (Pearce 2008 as cited in Baker, 2014). These adjectives contribute to the stereotype of associating women with judgments of attractiveness.

2. Allocational harms/Quality of service: This includes disparities in the quality of MT service across different gender groups. For example, post-editing studies carried out by Vanmassenhove et al. in 2018, indicate that feminine referents require disproportionately more manual correction than masculine ones in MT workflows, increasing cognitive and economic costs for the translation process.

3. Linguistic encoding of gender: genderless languages often generate artificial gender assignments in MT translations to gendered forms when translations, while for languages with a system of morphosyntactic agreement for MT translation defaults to masculine gender can result in misrepresentation.

4.Social gender connotations manifested through semantic derogation - a tendency for feminine forms to be subject to negative connotations in certain cases (such as the French words *couturier* (*fashion designer*) vs. *couturière* (*seamstress*)

or the English *master* vs. *mistress*); And epicene nouns such as the French word *médecin* (*doctor*) which is grammatically masculine but gender-neutral in meaning. When MT systems output *he* for *médecin*, they may reflect stereotype-driven gender assignment rather than the word's inclusivity.

5. Gender and language use (Linguistic features): Literature in the field of gender studies indicates differences in language use between genders, such as the use of hedging strategies, first-person pronouns, and prosodic exclamations. Studies by Lakoff (1975) suggest that there is a tendency for women to use more hedges (*sort of*, *maybe*) as a politeness strategy, though later corpus studies (Holmes 1990) show hedging is rather context-dependent than purely gendered. According to Argamon et al. (2003) women tend to use first-person singular pronouns (*I*) more often in personal narratives, while men use more abstract or third-person references, reflecting different discourse styles.

Recognition of gender bias has prompted some mitigation efforts from MT providers. In 2020, Google introduced upgrades to their gender-specific translation approach, acknowledging that machine learning models "can be skewed by societal biases reflected in their training data" (Johnson, 2020). They developed a three-step approach: (1) generate default translation; (2) if gendered, rewrite to alternative gender; (3) check both versions for accuracy. This resulted in Google Translate offering users both masculine and feminine translation options for ambiguous source sentences. While this represents progress, it addresses only a subset of gender bias issues. The approach works for single-sentence translations with clear occupational nouns but does not resolve problems in connected discourse, agreement chains, or contexts where gender is signalled through adjectives, participles, or pronouns rather than nouns. Costa-Jussà and de Jorge (2020) explored fine-tuning neural MT on gender-balanced datasets as an alternative approach, demonstrating improvements in specific language pairs.

Despite this substantial body of work, two limitations constrain current understanding and motivate the present study. First, most studies focus on English as either source or target language, limiting insights into how bias operates in direct translations between non-English pairs (Prates et al., 2020; Stanovsky et al. 2019). This English-centricity may obscure bias patterns specific to other linguistic contexts. Moreover, existing research relies predominantly on corpus analysis, which captures naturally occurring usage but cannot systematically test specific linguistic phenomena in controlled conditions.

Second, while Stanovsky et al. (2019) examined grammatical-gender languages and Prates et al. (2020) included gender-neutral sources, there are limited studies to systematically compare how the same gender-ambiguous content is rendered across typologically diverse target languages within a single controlled framework. Understanding whether bias manifests similarly or differently across language types, and whether typological features predict bias patterns, remains underexplored.

Additionally, architectural comparison is limited: Research has predominantly evaluated Google Translate's neural MT architecture. The rapid

deployment of large language models like ChatGPT, which approach translation as conversational text generation rather than statistical alignment, raises questions about whether architectural differences affect bias patterns. Whether conversational AI systems reproduce the same biases as translation-specific models, or whether their design enables greater flexibility in handling gender ambiguity, has received limited systematic investigation. This study addresses these gaps through controlled cross-linguistic comparison. By comparing two MT architectures (Google Translate's translation-specific neural MT versus ChatGPT's generative large language model) the study asks:

1. Do documented patterns of occupational stereotyping hold across different languages when tested systematically rather than through corpus analysis?
2. How do MT systems handle gender-ambiguity when translating into languages with different gender-marking systems, and do typological differences predict bias manifestation patterns?
3. Does architectural design affect gender bias: do Google Translate's neural MT and ChatGPT's conversational LLM exhibit different default patterns, or do both systems reproduce similar biases despite structural differences?

The present study does not operationalise these theoretical perspectives, but cites them as interpretive context for understanding why gendered MT outputs have attracted scholarly attention.

3 Methodology

This study adopts a challenge-set methodology (cf. Isabelle, Cherry, & Foster 2017), in which constructed examples are designed to test the translation systems of Google and ChatGPT in contexts where gender bias is known to manifest. This design aligns with established guidance in applied linguistics research, which emphasises coherence between research questions, data construction, and the scope of warranted claims in descriptive studies (Popescu, 2025). These examples were created following parameters identified in prior research on gender bias in MT (Prates, Avelar, & Lamb, 2020; Vanmassenhove, Hardmeier, & Way, 2018; Savoldi et al., 2022). The challenge-set approach enables controlled, systematic testing of specific phenomena: occupational nouns, ambiguous pronouns, adjective-gender associations, explicit gender agreement, epicene nouns, social gender connotation, and source language ambiguity, allowing for replicable evaluation and direct comparison across systems and language pairs.

The study comprises 175 test sentences organised into seven parameters (25 per parameter) which were translated by Google Translate and ChatGPT-4. The source languages were English for parameters 1-6 and Turkish for parameter 7. (see **Error! Reference source not found.**). The selected target languages were Romanian, French, Spanish, German and English (for parameter 7 only). The evaluation was done on the total number of 1.450 translations split onto the two systems.

Table 1. Test matrix for detecting gender bias in MT

Parameter	Subcategories and examples	Target Phenomenon
1. Occupational stereotyping	Male-coded: engineer, architect Female-coded: nurse, kindergarten teacher, florist Neutral/mixed: manager, writer	Stereotype-based gender assignment across occupation types
2. Pronoun ambiguity	Singular *they*: *The student said they would arrive later* Ambiguous *you*: *You are welcome to join us* Forced-gender verb constructions: *You are tired after the long journey* Ambiguous antecedents: *The teacher spoke to the parent before they left* Epicene nouns + pronouns: *The colleague said they would help*	Resolution strategies for five types of pronoun ambiguity
3. Adjective Descriptions	Attractiveness: beautiful, graceful Strength/toughness: tough, assertive Competence: intelligent, skilled Emotional/caring: caring, sensitive Professional/neutral: professional, experienced	Gender assignment based on adjective social connotations with unspecified subjects
4. Grammatical Gender Agreement	Feminine pronouns + occupations: *The engineer explained her design process* Masculine pronouns + occupations: *The nurse explained his daily routine* Contextual cues across clauses: *The chef prepared the meal. She used fresh ingredients*	Maintenance of explicit gender cues; stereotype override detection
5. Epicene Nouns	Family roles: parent, cousin Professional roles: teacher, doctor Social roles: neighbour, citizen	Treatment of inherently gender-neutral vocabulary across role types
6. Social gender Connotation	Prestige term pairs: designer/fashion designer/costume designer; professional/medical professional/legal professional	Semantic derogation testing; gender assignment for high-status terms
7. Ambiguous Context (Turkish Source)	Gender-neutral names (e.g. Deniz, Özgür) tested across: Achievement: *Ege won the competition. He/She is hardworking* Occupation: *Deniz is a teacher. He/She loves his/her job* Family context: *Derin takes care of his/her child* Personal context: *Nehir loves cooking* Professional context: *Irmak motivated the team*	Gender assignment from radically neutral source (Turkish: no grammatical gender, single pronoun *o*)

Sentences were organised by parameter and submitted to Google Translate (neural machine translation architecture) and ChatGPT-4 (large language model, conversational AI architecture) via web interfaces. Test sentences were submitted in batches organised by parameter category. Parameters 1, 3, 5, and 6, comprising independent sentences with no contextual interdependencies, were batched in groups of 15-25 sentences per prompt. Parameters 2, 4, and 7, containing contextually sensitive constructions where sequential presentation might influence system behaviour, were submitted in smaller batches of 5-10 sentences to minimise pattern recognition effects that could artificially alter outputs. Each translation was coded according to a framework adapted from Savoldi et al. (2021) with modifications to capture observed phenomena. (see **Error! Reference source not found.**)

Table 2. Coding scheme

CODE	DEFINITION	EXAMPLE
M-DEF	Masculine default	Gender-neutral source: masculine target without contextual justification (e.g., EN *colleague* → RO *colegul* masc.)
F-DEF	Feminine default	Gender-neutral source: feminine target without contextual justification (e.g., EN *secretary*→ RO *secretara* fem.)
BOTH	Both options provided	System offers masculine and feminine alternatives (e.g., FR *fatigué(e)*, ES *orgulloso/a*, EN he*/she*)
GENNEU	Gender neutral	Neutrality maintained through target language structures (e.g., German neuter *das Teammitglied*)
GENAVOID	Gender avoidance	System employs rephrasing, demonstrative pronouns, formal plural constructions, or alternative grammatical resources
CORRECT	Accurate maintenance	Explicit source gender maintained correctly. Code used in parameter 4 only.
ERR	Error	Grammatical invention/contradiction (e.g. *cetățeanul/cetățeanul* → identical forms provided for both genders)

Coding was performed through systematic review of all 1.450 translations. For each parameter, translations were evaluated against expected outcomes defined in the test matrix (**Error! Reference source not found.**). The analysis employed both quantitative aggregation and qualitative linguistic examination. Quantitative analysis calculated frequency distributions of each code by parameter, target language, and system, enabling identification of systematic patterns. Qualitative analysis examined specific translation outputs to identify mechanisms underlying bias patterns.

This study reports observed outputs for a constructed challenge set. Coding was performed by a single analyst and no inter-annotator agreement was calculated. The analysis is descriptive and does not aim at statistical inference beyond the tested items. Because MT and LLM systems change over time, findings reflect outputs observed at the time of testing under the stated prompt batching procedure.

4 Results

4.1 Parameter 1. Occupational stereotyping

Of 200 translations (25 sentences × 4 languages × 2 systems), 158 (79%) defaulted to masculine, distributed evenly between ChatGPT (83 sentences) and Google Translate (75 sentences). Feminine defaults occurred in 34 instances (17%); both gender options in only 6 cases (3%, all French ChatGPT). Spanish exhibited highest masculine defaults (44 sentences); Romanian and German highest number of feminine defaults (12 and 10 instances respectively). French alone offered both feminine and masculine options in translation, though in under 5% of tested cases.
Traditionally male-coded professions (*engineer, surgeon, architect, pilot, CEO, professor*) produced 100% masculine defaults across all languages and both systems, with no feminine alternatives or neutral formulations. This uniformity across indicates occupational prestige correlates perfectly with masculine assignment regardless of system design. Traditionally female-coded professions (nurse, kindergarten teacher, housekeeper, librarian, receptionist, florist, babysitter) showed

59% feminine defaults (33 of 56 sentences), concentrated in Romanian (12 instances). Spanish diverged notably: 8 masculine defaults for female-coded professions (7 from ChatGPT), including *el bibliotecario (librarian)*, *el enfermero (nurse)*, *el niñero (babysitter)*. Neutral professions (*manager, writer, lawyer, photographer, designer, journalist, consultant*) produced 97% masculine defaults (85 of 88). Only two French ChatGPT translations offered both options: *l'assistant(e)*, *le directeur/la directrice*. This near-total masculine default indicates that, for the tested items, the systems frequently select masculine forms when source gender is unspecified.

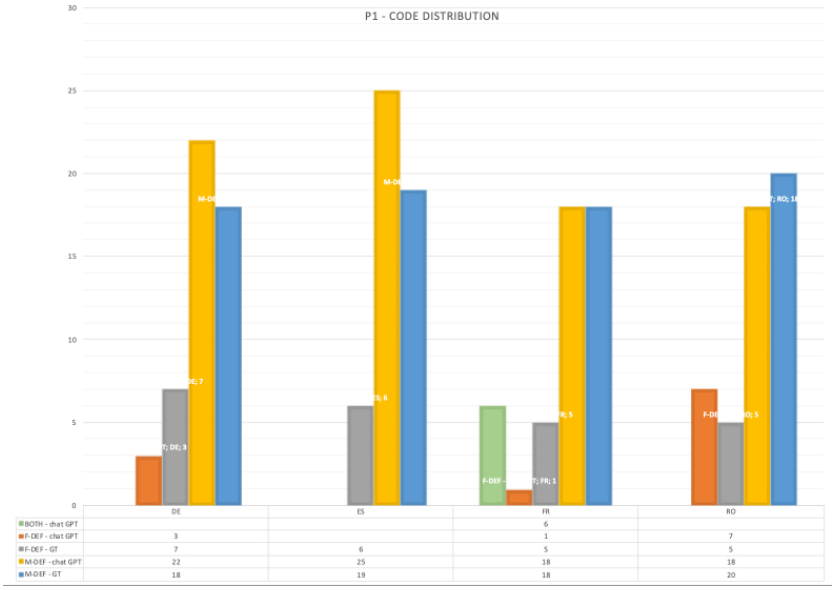

Figure 1. Occupation stereotyping code distribution chart

ChatGPT provided all 6 both-gender instances, suggesting greater meta-awareness, though capability was rarely deployed. Google Translate showed mechanical consistency, prioritising single outputs without alternatives. Cross-linguistically, Spanish showed highest masculine rates; French unique both-gender provision and Romanian/German retained elevated feminine for care professions but defaulted masculine otherwise.

4.2 Parameter 2. Pronoun ambiguity

Across 200 translations split into five subcategories (singular they, ambiguous you, forced-gender verb constructions, ambiguous antecedents, and epicene nouns with pronouns)138 outputs (69%) defaulted to masculine, split almost evenly between Google Translate (70) and ChatGPT (68). Both-gender options appeared in 28 cases (14%), overwhelmingly from ChatGPT (25). Gender-avoidance strategies occurred in 13 instances (6.5%), mostly from Google Translate (8). Romanian showed the

highest masculine default rate (43/138 instances, 31%), German the lowest (24 instances, 18%).

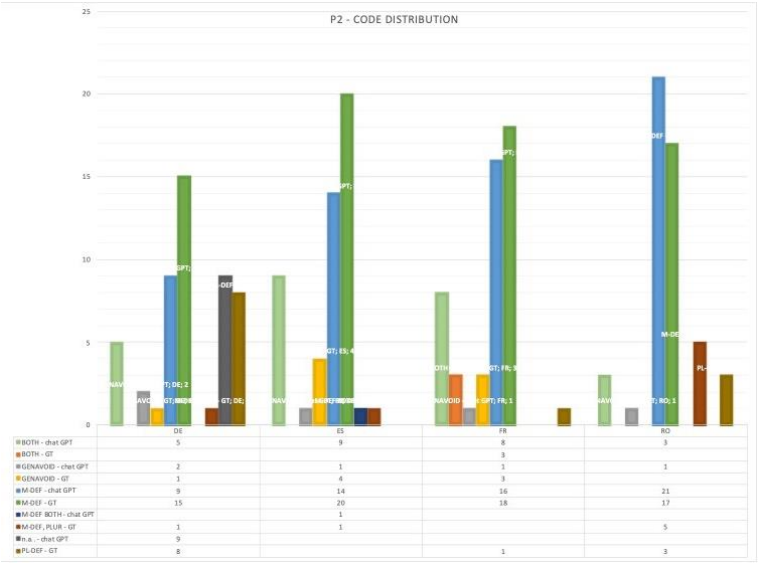

Figure 2. Pronoun ambiguity results overview

Masculine defaults varied by subcategory, revealing a hierarchy of bias. Epicene nouns paired with pronouns produced 39 masculine outputs out of 40 (97.5%) across all languages; only one Spanish translation from Google Translate avoided gendering. Ambiguous antecedents yielded 38 masculine defaults (95%), with only two avoidance cases (both ChatGPT in German) and no both-gender options, indicating that referential complexity intensifies masculine resolution. Singular they resulted in 34 masculine defaults (85%). German was the sole exception: ChatGPT offered both masculine and feminine alternatives in five cases. All other languages assigned a single, predominantly masculine form, erasing the source's intended neutrality. Forced-gender verb constructions showed a different pattern: 18 masculine defaults (45%), mostly from Google Translate (13). Romanian accounted for 10 of these. Ten translations (9 from ChatGPT) provided both-gender options, exclusively in French and Spanish, using inclusive typographic conventions (e.g., *orgulloso/a, fatigué(e)*). German produced no masculine defaults due to its gender-neutral participial structures. Ambiguous *you* pronouns generated 13 both-gender outputs (11 ChatGPT, 2 Google Translate), unevenly distributed across languages (French 6, Spanish 4, Romanian 3, German 0). Gender avoidance appeared in 7 cases, concentrated in Spanish.

System comparison shows that ChatGPT acknowledges ambiguity more frequently but resolves it similarly to Google Translate. ChatGPT produced both-gender options eight times more often (25 vs. 3 sentences) and consistently appended meta-textual invitations to adjust gender. Yet when producing single-gender outputs, both systems defaulted to masculine at nearly identical rates

(ChatGPT 34%, Google Translate 35%). Google Translate used avoidance slightly more often (7 vs. 2). ChatGPT's alternatives clustered in specific contexts (forced-gender verbs in French/Spanish and singular they in German) suggesting targeted rather than systematic mitigation. Cross-linguistic patterns parallel those observed in occupational stereotyping. French showed the most both-gender options, German uniquely offered alternatives for singular they and used avoidance for ambiguous antecedents, though masculine defaults remained frequent (24). Spanish showed the highest avoidance rate while Romanian produced the most masculine defaults overall.

Pronoun-ambiguity results reveal deeper structural bias than occupational stereotyping. When the source explicitly encodes neutrality, through epicene nouns or singular *they*, systems overwhelmingly erase it: 97.5% masculine defaults for epicene nouns and 85% for singular *they*. Across the tested subcategories, masculine defaults were most frequent for epicene nouns with pronouns and ambiguous antecedents, and less frequent for forced-gender verb constructions. This inverse relationship suggests that MT models interpret gender absence as a gap requiring masculine resolution. ChatGPT's meta-awareness operates retrospectively: it offers alternatives only after producing a masculine default, placing the choice of correction on users.

4.3 Parameter 3. Adjective descriptions

Translation of adjective-subject pairs (e.g., *The speaker was attractive*, *The leader was tough*) tested whether systems assign gender based on social connotations rather than linguistic necessity. Across 200 translations, 192 (96%) defaulted to masculine, split evenly between ChatGPT (98) and Google Translate (94), indicating highly consistent behaviour across systems. Attractiveness adjectives (*beautiful, attractive, elegant, charming, graceful*) produced masculine forms in 36/40 cases (90%). The only feminine outputs came from Google Translate for *The artist was beautiful* in German (*Die Künstlerin war schön*) and Spanish (*La artista era hermosa*). Two Romanian gender-avoidance cases were the only other deviations from the dominant pattern. Strength-related adjectives (*tough, strong, assertive, powerful, determined*) showed even stronger masculine bias: 38/40 translations (95%). The two exceptions appeared in German Google Translate outputs using gender-neutral professional terms (*Die Führungskraft war …*), illustrating that German offers neutral alternatives that systems deployed rarely and inconsistently throughout the tested sentences. Competence adjectives (*intelligent, skilled, competent, talented, capable*) yielded 40/40 masculine defaults across all languages and systems, with no feminine, neutral, or dual-gender forms. This uniformity suggests that competence, unlike attractiveness or strength, triggers categorical masculine assignment. Emotional and caring adjectives (*caring, sensitive, warm, gentle, nurturing*) also produced masculine defaults in 39/40 cases (97%). The sole exception again came from German Google Translate (*Die Pflegekraft war fürsorglich*), using the

gender-neutral *Pflegekraft*. Despite feminine stereotypes associated with care work, systems overwhelmingly imposed masculine forms. Professional and neutral adjectives (*professional, experienced, organised, creative, dedicated*) followed the same pattern: 39/40 masculine defaults, with one German neutral form (*Das Teammitglied war organisiert*).

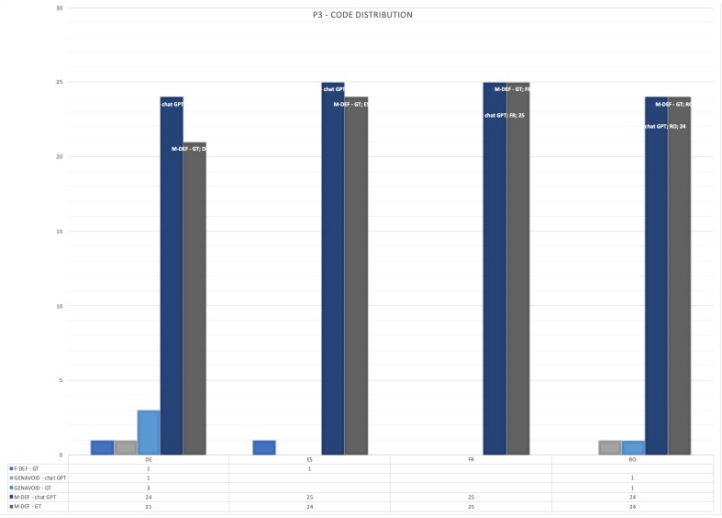

Figure 3. Adjective descriptions results overview

Overall, systems assign gender in adjective-subject constructions primarily through grammatical obligation rather than adjective semantics. Even adjectives stereotypically associated with femininity trigger masculine defaults when subjects are unspecified professionals. Gender-neutral alternatives appear only in German and only via Google Translate, accounting for less than 2% of outputs. Cross-linguistic consistency was high as well: Romanian, French, Spanish, and German all show roughly 96% masculine defaults, with German alone producing neutral alternatives. This uniformity suggests that the observed behaviour is not fully explained by target-language grammatical constraints alone.

4.4 Parameter 4. Grammatical gender agreement

Parameter 4 tested whether systems maintain explicitly marked gender through pronouns like *her research* or *his method.* The overall accuracy of 84% initially suggests reasonable performance, but 31 failures reveal stereotypes overriding explicit cues. ChatGPT performed moderately better (94 correct translations vs. 75 from GT) with German showing the highest accuracy (67 correct translations) while Romanian had the lowest (34). This cross-linguistic variation suggests that certain language structures facilitate accurate gender transfer while others create vulnerability to bias-driven overrides. The most striking failures occurred when explicit gender cues contradicted occupational stereotypes. 13 translations of

sentences containing masculine pronouns with traditionally female-coded professions (*The secretary organised his boss's schedule*, *The receptionist answered his phone calls*, *The kindergarten teacher read his students stories*), were rendered in feminine forms across all target languages, with Google Translate responsible for 10 of these. Systems ignored masculine *his* and defaulted to feminine based on occupation alone. The inverse pattern appeared with feminine pronouns paired with traditionally male-coded professions. 6 Romanian translations, representing ¾ of applicable Romanian sentences, converted feminine pronouns to masculine despite explicit *her* cues: *The engineer explained her design process*, *The architect showed her building plans*, *The professor published her latest book*. Google Translate produced 5 of these errors; ChatGPT only one. However, examination of specific outputs reveals different error types. For *The architect showed her building plans*, ChatGPT's Romanian translation rendered *Arhitecta și-a prezentat planurile clădirii* (*The architect[fem.] presented the building's plans*), maintaining gender accuracy, but Google Translate produced *Arhitectul i-a arătat planurile de construcție* (*The [masculine] architect showed her [someone else] the construction plans*), simultaneously changing the architect's gender from feminine to masculine and introducing a phantom third party. ChatGPT rendered *The engineer explained her design process* correctly in French, Spanish, and German, suggesting language-specific processing differences where Romanian proved most vulnerable to masculine override. Spanish produced five feminine defaults for masculine pronoun sentences despite explicit *his*. This aligns with Parameter 1 patterns, indicating bidirectional instability where stereotypes override syntax unpredictably.

Contextual gender cues in sentences where gender appears in a second clause (*The chef prepared the meal. She used fresh ingredients*) revealed sharper system divergence. ChatGPT maintained gender consistency across both sentences with complete accuracy, while Google Translate failed in 42% of cases. Google Translate typically defaulted the first sentence to masculine based on occupation and either used a neutral phrase through the second sentence (Romanian: *Bucătarul... A folosit*; Spanish: El chef... Usó) or created internal contradiction (French: *Le chef... Elle a utilisé*). The example from French (masculine noun with feminine pronoun) indicates Google Translate processes sentences independently without maintaining referential coherence across clauses. The sole parameter subset with perfect accuracy involved sentences with explicitly gendered nouns: *The actress received her award gracefully*, *The businessman presented his quarterly results.* When both noun and pronoun carried gender marking, systems did not fail. This indicates that errors emerge specifically when occupational terms are gender-neutral or epicene in English, requiring systems to rely on pronoun cues alone, precisely the context where stereotypical associations interfere.

In the tested items, explicit pronoun cues were not always preserved in output, particularly when paired with stereotyped occupational terms. Systems override pronouns when contradicting stereotypes. The identified errors were asymmetric: Romanian translations were vulnerable to masculine override; Spanish to feminine, while German was the most resistant. ChatGPT's superior contextual

performance suggests architectural advantages in referential coherence, though both systems failed when stereotypes conflicted with syntax.

4.5 Parameter 5. Epicene nouns

Epicene nouns assess whether MT systems preserve source-language neutrality when English provides explicitly non-gendered role terms. Across 200 translations, systems overwhelmingly imposed gender: 163 outputs (81.5%) defaulted to masculine**,** distributed evenly across languages and systems. Only 21 translations (10.5%) maintained gender-neutral formulations, and 5 (2.5%) employed gender-avoidance strategies. Two feminine defaults appeared exclusively in German Google Translate outputs. Romanian produced the highest number of both masculine and feminine forms (5 of the 7 total), indicating that dual-gender rendering is possible but rarely activated.

French accounted for the largest share of neutral outputs **(**12/21), split evenly between systems, reflecting the availability of genuinely epicene lexical items (*le parent, l'enfant*). German and Spanish produced fewer neutral forms, though German occasionally used neuter compounds (*das Geschwisterkind, die Führungskraft*). Romanian showed masculine defaults dominating across all subcategories. Two default to plural outputs appeared in family-role contexts (DE *Die Eltern* for *the parent* and RO *Descendenții* for *the offpring*). While statistically marginal, these plural defaults are noteworthy: they show that systems sometimes circumvent gender assignment by shifting number rather than preserving neutrality, a strategy that avoids gender marking without explicitly encoding ambiguity. Masculine defaults were nearly identical across systems (ChatGPT 83; Google Translate 80), indicating that both models resolve epicene nouns in similar ways despite architectural differences. Neutral outputs were also balanced (ChatGPT 10; Google Translate 11). Gender-avoidance strategies appeared slightly more often in Google Translate (3 vs. 2), though the difference is minimal. The two feminine defaults occurred only in German Google Translate outputs, suggesting isolated lexical choices rather than systematic behaviour.

Cross-linguistic consistency, excluding Romanian's small cluster of both masculine and feminine forms, indicates that, in this case, bias arises from shared training data patterns and model architectures rather than from target-language grammatical constraints. French's comparatively high neutral rate reflects lexical affordances, while German's occasional neuter compounds show that ambiguity can be preserved when the lexicon permits it. However, these exceptions do not alter the overall pattern: English gender-neutral vocabulary becomes masculinised in translation in more than four out of five cases**.**

4.6 Parameter 6. Social gender connotation

Parameter 6 tested whether systems assign gender differently based on occupational prestige, examining terms like *designer*, *chef*, *artist*, *professional*, and *expert* paired with evaluative adjectives. Of 200 translations, 181 defaulted to masculine (90%), distributed nearly evenly between ChatGPT (94) and Google Translate (87). 16 defaults appeared, with 13 from Google Translate. The distribution of feminine defaults was concentrated: 15 of 16 appeared in translations of three sentences containing *secretary* (*The administrative secretary coordinated meetings, The executive secretary handled confidential matters, The secretary managed office communications)* across all target languages. The single remaining feminine default appeared for *costume designer* in German. This pattern aligns with parameters 1 and 4 where systems assigned gender based on occupational stereotypes regardless of source language neutrality. However, the absence of semantic derogation, where high-prestige terms would be translated as lower-prestige equivalents when feminine, suggests that bias operated primarily through gender assignment rather than through translating *designer* as *seamstress* or *chef* as *cook*. Systems maintain lexical equivalence while imposing gendered forms. The near-complete masculine default for prestige professional terms confirms that high-status professional identity correlates with masculine gender assignment. ChatGPT and Google Translate performed virtually identically, indicating shared training data patterns where professional prestige and masculine grammatical gender co-occur statistically.

4.7 Parameter 7. Ambiguous context

Parameter 7 examined translation from Turkish, a language with no grammatical gender and a single third-person pronoun *o*, into five target languages (Romanian, German, French, Spanish and English which was used as the source language for the previous parameters). Turkish personal names such as *Deniz, Özgür, Evren,* are gender-neutral, creating ambiguity through the absence of grammatical cues, gendered morphology, and with minimal stereotypical associations. This parameter therefore tests how systems assign gender when the source language provides no justification for doing so.

The results differ sharply from parameters 1-6. Masculine defaults declined to 130/250 (52%)**,** while 76 translations (36%) employed feminine forms, by far the highest feminine rate in the study. (see **Error! Reference source not found.**) Systems also produced 23 gender-neutral outputs and 21 versions with both masculine and feminine, the latter representing the second-highest rate across all parameters. The reduction of masculine dominance indicates that when gender cues are entirely absent, systems distribute gender more variably, though masculine remains the majority choice.

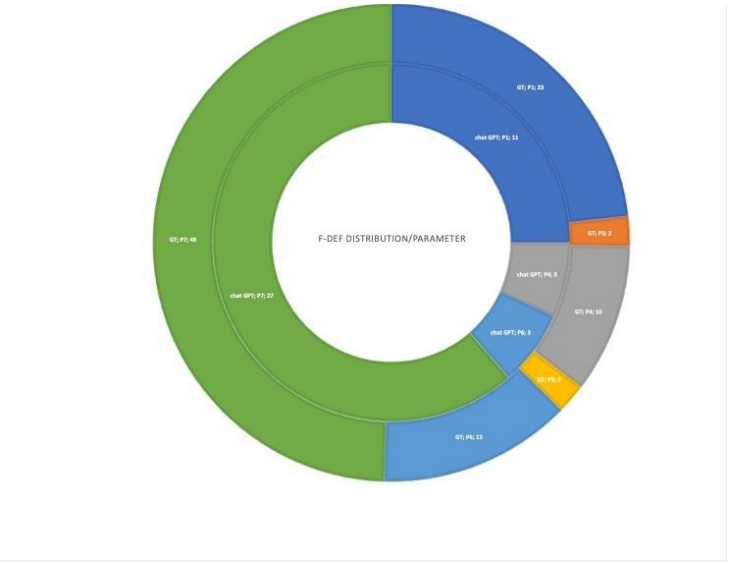

Figure 4. Overall default to feminine distribution / test parameter

ChatGPT produced all 21 dual pronoun outputs**,** concentrated in English (10), French (10), and German (1). Google Translate never offered alternatives, but generated substantially more feminine defaults (45 vs. 27) and more gender-neutral forms (21 vs. 8), especially in Spanish (18 of 29 outputs). The patterns indicated distinct strategies: ChatGPT signalled ambiguity through dual forms in select languages, whereas Google Translate resolved ambiguity through a wider distribution of feminine and neutral assignments.

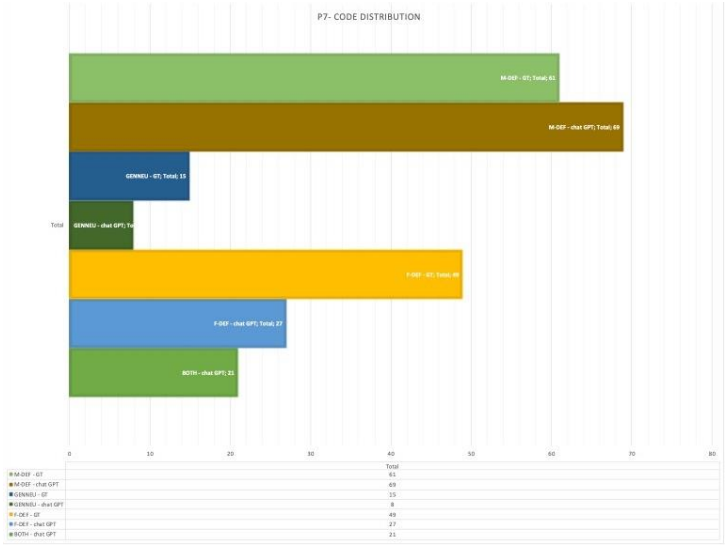

Figure 5. Ambiguous context results distribution

Subset analysis shows that context strongly shapes gender assignment: Achievement contexts produced mostly masculine defaults (30), with feminine forms clustering around names like *Selin*, which, though formally neutral, may skew feminine in training data. Variation across languages suggests name-frequency effects rather than systematic rules. Educational contexts showed similar patterns: feminine defaults concentrated around *Naz*, another officially neutral but commonly feminine name, indicating that systems draw on statistical name-gender associations. Family contexts were the only subset where feminine exceeded masculine (15 vs. 11). Sentences involving caregiving (*Derin takes care of his/her child*) triggered feminine defaults across nearly all languages and systems, suggesting that caregiving contexts activate stereotypically feminine associations even when names are neutral. Occupational contexts revealed the strongest system divergence. Google Translate produced 14 feminine defaults for traditionally female-coded professions (nurse, teacher, lawyer) across all languages, while ChatGPT offered 10 masculine and feminine options (English/French) and otherwise defaulted masculine. This reverses earlier parameters, where Google Translate was more rigidly masculine. One likely explanation is training-data frequency: Turkish-to-target corpora may associate these professions with feminine forms. Personal-activity contexts (e.g., cooking, hobbies) showed mixed patterns: 30 masculine, 11 neutral (mostly Spanish), and 7 feminine defaults, with *Nehir + cooking* producing most feminine outputs. Professional-leadership contexts unexpectedly produced 19 feminine defaults, reversing the masculine leadership bias seen in Parameters 1-3.

Overall, Turkish revealed that when gender cues are entirely absent, MT systems behave less uniformly, but masculine remains the dominant default, shaped by name frequency, contextual stereotypes, and system-specific training distributions.

5 Conclusion

This study examined gender resolution patterns in Google Translate and ChatGPT through a descriptive evaluation of 1,450 translations across seven parameters and five typologically diverse target languages. Three research questions guided the investigation.

The first question examined whether occupational stereotyping patterns reported in prior work recur systematically across languages when tested using a controlled challenge-set. Within the tested items, traditionally male-coded professions were rendered with masculine forms in all observed cases across the four target languages and both systems, while neutral professions showed predominantly masculine outputs. This consistency indicates that, for the evaluated sentence set, masculine forms were frequently selected when occupational gender was not specified in the source.

The second question addressed how the systems handle gender ambiguity across languages with different gender-marking systems and whether typological differences are associated with variation in output. Across the tested parameters,

source-language neutrality was frequently not preserved in translation: epicene nouns, singular *they*, and ambiguous antecedents were rendered with masculine forms in the majority of cases. Differences across languages were observed at the margins, with French producing a higher number of dual-gender forms, German occasionally offering alternatives for singular *they*, and Spanish employing gender-avoidance strategies more frequently than other languages. These variations suggest that typological features may influence how gender is realised in output, but they do not eliminate the predominance of masculine forms within the tested data.

The third question explored whether differences between the two systems were observable in gender resolution behaviour. When producing single-gender outputs, both systems showed similar distributions of masculine defaults across several parameters. ChatGPT more frequently provided alternative gendered renderings or explicit invitations to adjust gender, while Google Translate typically produced a single form. These differences indicate variation in how ambiguity is acknowledged at the interface level, but they do not substantially alter the distribution of primary gendered outputs within the tested set. An exception was observed in sentences requiring cross-clausal reference, where ChatGPT more consistently preserved gender agreement across clauses than Google Translate in the evaluated examples.

Across parameters, the results document recurrent patterns in how gender is realised in MT output, including the frequent selection of masculine forms in occupational contexts, limited preservation of source-language neutrality, and variability associated with source language and target-language grammatical resources. In cases where explicit pronoun cues conflicted with occupational stereotypes, these cues were not always maintained in the output, indicating that grammatical and lexical choices may interact in non-uniform ways within the tested translations.

The study's limitations include the size and constructed nature of the test set, single-date testing of systems that are subject to change, and a focus on binary gender distinctions. Accordingly, the findings are descriptive and restricted to the observed outputs under the stated conditions. Future research could extend this work through larger corpus-based evaluations, examination of gender distributions in parallel training data, inclusion of additional language families, and systematic investigation of non-binary gender representation in MT.

Despite recent mitigation efforts by system developers, the observed patterns indicate that gender resolution in MT remains uneven across languages and contexts. For translators and language professionals, the findings underscore the importance of critically assessing MT output in gender-sensitive contexts. More broadly, the study contributes a structured empirical record of gendered output patterns that may inform further methodological, linguistic, and applied research on gender in machine translation.

References

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style. *Text & Talk, 23*(3), 321–346. https://doi.org/10.1515/text.2003.014

Bab.la. (n.d.). Gender bias. In *Bab.la dictionary*. Retrieved May 11, 2024, from https://en.bab.la/dictionary/english/gender-bias

Baker, M. (2006). *Translation and conflict: A narrative account*. Routledge. https://doi.org/10.4324/9780203099919

Baker, P. (2014). *Using corpora to analyse gender*. Bloomsbury.

Bentivogli, L., Savoldi, B., Negri, M., Di Gangi, M. A., Cattoni, R., & Turchi, M. (2020). Gender in danger? Evaluating speech translation technology on the MuST-SHE corpus. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 6923–6933). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.619.pdf

Butler, J. (1999). *Gender trouble: Feminism and the subversion of identity*. Routledge.

Comrie, B. (1999). Grammatical gender systems: A linguist's assessment. *Journal of Psycholinguistic Research, 28*, 457–466. https://doi.org/10.1023/A:1023212225540

Corbett, G. G. (Ed.). (2014). *The expression of gender*. De Gruyter Mouton.

Costa-Jussà, M. R., & de Jorge, A. (2020). Fine-tuning neural machine translation on gender-balanced datasets. In M. R. Costa-Jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing* (pp. 26–34). Association for Computational Linguistics. https://aclanthology.org/2020.gebnlp-1.3

Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

Ghosh, S., & Caliskan, A. (2023). ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across Bengali and five other low-resource languages. In F. Rossi, S. Das, J. Davis, K. Firth-Butterfield, & A. John (Eds.), *Proceedings of the 2023 AAAI/ACM conference on AI, ethics, and society (AIES '23)* (pp. 901-912). Association for Computing Machinery. https://doi.org/10.1145/3600211.3604672

Gygax, P. M., Elmiger, D., Zufferey, S., Garnham, A., Sczesny, S., von Stockhausen, L., Braun, F., & Oakhill, J. (2019). A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in Psychology, 10*, Article 1604. https://doi.org/10.3389/fpsyg.2019.01604

Hellinger, M., & Bussmann, H. (Eds.). (2001). *Gender across languages: The linguistic representation of women and men* (Vol. 1). John Benjamins. https://doi.org/10.1075/impact.9

Holmes, J. (1990). Hedges and boosters in women's and men's speech. *Language & Communication, 10*(3), 185–205. https://doi.org/10.1016/0271-5309(90)90002-S

Isabelle, P., Cherry, C., & Foster, G. (2017). A challenge set approach to evaluating machine translation. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2486–2496). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1263

Jakobson, R. (1972). Verbal communication. *Scientific American, 227*(3), 72–81.

Johnson, M. (2020, April 22). A scalable approach to reducing gender bias in Google Translate. *Google Research*. https://research.google/blog/a-scalable-approach-to-reducing-gender-bias-in-google-translate/

Kayser-Bril, N. (2020, September 17). Female historians and male nurses do not exist, Google Translate tells its European users. *AlgorithmWatch*. https://algorithmwatch.org/en/google-translate-gender-bias

Kramer, R. (2014). Gender in Amharic: A morphosyntactic approach to natural and grammatical gender. *Language Sciences, 43*, 102–115. https://doi.org/10.1016/j.langsci.2013.10.004

Lakoff, R. (1975). *Language and women's place*. Harper & Row.

Nedlund, E. (2019, November 12). Apple Card is accused of gender bias. Here's how that can happen. *CNN Business*. https://edition.cnn.com/2019/11/12/business/apple-card-gender-bias/index.html

Popescu, T. (2025). *Research in applied linguistics and language education: Design, methods, and analysis*. Presa Universitară Clujeană. https://doi.org/10.29302/ResearchApplLing_LangEduc.popescu.t

Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications, 32*, 6363–6381. https://doi.org/10.1007/s00521-019-04144-6

Robert F. Kennedy Human Rights. (2023). How women shaped the Universal Declaration of Human Rights. https://rfkhumanrights.org/our-voices/how-women-shaped-the-universal-declaration-of-human-rights-2/

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics, 9*, 845–874. https://doi.org/10.1162/tacl_a_00401

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2022). Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In S. Mureșan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of*

the Association for Computational Linguistics (pp. 1807–1824). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.127

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 1679–1684). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1164

United Nations. (2019). Women who shaped the Universal Declaration of Human Rights. https://www.un.org/en/observances/human-rights-day/women-who-shaped-the-universal-declaration

Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). Getting gender right in neural machine translation. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in Natural Language Processing* (pp. 3003–3008). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1334