A Chinese Challenge Set to Assess Gender Bias in Automated Translation

Xiaolan Xu[1], Sara Mendes[1, 2], Yu-Yin Hsu[3]

[1] Faculty of Arts and Humanities, University of Lisbon
xuxiaolan@edu.ulisboa.pt, s.mendes@edu.ulisboa.pt
[2] Center of Linguistics of the University of Lisbon
[3] Department of Language Science and Technology, The Hong Kong Polytechnic University
yyhsu@polyu.edu.hk

Abstract

Gender bias in automated translation (AT) is well-documented for English-source language pairs, while source languages generally lacking grammatical gender remain largely understudied. This paper addresses the Chinese–Portuguese direction, a language pair that has received little attention in this context. We developed a Chinese challenge set of 495 sentences constructed from 45 occupations across an eleven-sentence template matrix, systematically varying the type and syntactic position of gender cues: no cue, explicit prenominal modifiers, and coreferential pronouns varying in position and syntactic complexity. We tested two commercial neural machine translation (NMT) systems and six large language models (LLMs) with this challenge set. Results show a clear hierarchy of cue effectiveness: explicit prenominal modifiers yield universal 100% accuracy; in the absence of gender cues, models predominantly default to masculine forms; and coreferential pronouns in complex sentences reveal a pronounced masculine–feminine asymmetry. Also, LLMs demonstrate more symmetric gender cue processing than NMT systems. Our data and code are available at https://github.com/xu-xiaolan/chinese-challenge-set-gender-bias.

1 Introduction

Automated translation (AT), encompassing both neural machine translation (NMT) systems and general-purpose large language models (LLMs) used for translation tasks, has become a central tool for cross-linguistic communication, although its reliance on large-scale naturally occurring data makes it susceptible to reproducing and amplifying the biases present in human-generated text (Zhao et al., 2017; Blodgett et al., 2020). Gender bias is among the most thoroughly studied bias phenomena: training corpora tend to over-represent male-centric contexts in comparison to female-centric ones, and models trained on such data may encode stereotypes that surface systematically in their outputs (Savoldi et al., 2021).

In this context, the research on gender bias in MT has expanded substantially over the past decade. Stanovsky et al. (2019) introduced the first challenge set and evaluation protocol for measuring gender bias in MT, revealing systematic inaccuracies across four commercial systems and two academic models. Subsequent work has proposed a range of mitigation strategies (Font and Costa-Jussà, 2019; Saunders and Byrne, 2020; Costa-Jussà et al., 2022; Fleisig and Fellbaum, 2022). Challenge sets and benchmarks have been developed for numerous language pairs and systems (Piergentili et al., 2023; Sewunetie et al., 2024), and the role of sentence context in shaping AT gender behavior has received growing attention (Sharma et al., 2022; Hackenbuchner et al., 2024; Gkovedarou et al., 2025). Despite this breadth of research topics, the existing literature exhibits a consistent skew: English is almost invariably the source language, and the language pairs studied are largely Indo-European, leaving

typologically distinct languages such as Chinese–Portuguese largely unexamined.

Gender bias in AT manifests in several ways, including the default assignment of masculine forms to gender-ambiguous referents and the stereotypical gendering of occupational nouns (Bonifácio et al., 2025). These errors are particularly consequential when the target language encodes grammatical gender morphologically: in Portuguese, a single mistaken gender assignment propagates across determiners, adjectives, and participial forms through morphosyntactic agreement. Furthermore, the translation of gender information is particularly challenging in the Chinese–Portuguese direction, as in Chinese, gender is marked only in written forms of third-person pronouns (他/她/它, 'he/she/it', all pronounced *tā*), while nouns carry no morphological gender marking. In this way, when translating from Chinese into Portuguese, AT systems must resolve a gender value that is simply absent in the source.

Besides the linguistic asymmetry outlined above, studying the Chinese–Portuguese language pair is also motivated by several practical reasons. Chinese and Portuguese are among the most widely spoken languages globally, and expanding economic and cultural ties between China and Portuguese-speaking countries have substantially increased translation demand in this direction, generating a substantial and increasing translation demand between these two languages. Also, the Special Administrative Region of Macao, where both Chinese and Portuguese are the official languages and which China has been promoting and supporting as a hub between China and Portuguese-speaking countries, represents a unique context in which Chinese–Portuguese translation is an everyday institutional reality, making the accuracy and fairness of AT systems involving this language pair a matter of immediate practical consequence.

The present study addresses the construction and evaluation of a Chinese challenge set, designed to assess how different linguistic conditions in Chinese source sentences impact gender bias in Chinese–Portuguese AT outputs. The challenge set is developed with three objectives: (1) to develop a replicable methodology and dataset for evaluating gender bias in AT when Chinese is the source language; (2) to evaluate how different linguistic conditions impact gender bias across AT systems; and (3) to propose a replicable methodology for constructing gender bias challenge sets for AT when no such datasets are available.

To achieve these goals, we evaluate the outputs of eight AT systems: two commercial NMT engines and six recently released LLMs.

Section 2 reviews related work on gender bias in AT and provides the linguistic background for the study. Sections 3 and 4 detail the experimental setup and Chinese dataset construction. Section 5 reports the results, and Section 6 discusses the findings and presents conclusions and future work.

2 Related Work and Background

2.1 Gender Bias in AT

Gender bias in MT has attracted considerable attention over the past decade, since MT training data often reflects societal gender imbalances, leading models to perpetuate or amplify these biases (Zhao et al., 2017). Such inaccuracies not only distort meaning but also increase post-editing time and cost (Savoldi et al., 2024).

To evaluate and measure gender bias across MT systems, Stanovsky et al. (2019) established the foundational benchmark WinoMT, a challenge dataset and evaluation protocol, which revealed systematic inaccuracies in both industrial and academic systems. Subsequent work proposed various mitigation strategies for NMT models, including debiasing word embeddings (Font and Costa-Jussà, 2019), transfer learning (Saunders and Byrne, 2020), and novel learning frameworks (Fleisig and Fellbaum, 2022), though these efforts have focused predominantly on English as the source language. Even if work has been developed on other languages, e.g. Prates et al. (2020) demonstrated that gender-neutral languages can serve as a lens for examining gender bias in MT; and Cho et al. (2019) constructed a Korean gender bias test set, highlighting the need for evaluation resources for other languages, Chinese as a source language, and the Chinese–Portuguese language pair remain underexplored.

Recent work has begun to examine gender bias in AT outputs produced by LLMs, revealing that base LLMs exhibit a higher degree of gender bias than NMT models in English-to-Catalan and English-to-Spanish translation (Sant et al., 2024), although according to these authors, prompt engineering reduces gender bias by up to 12%. Additionally, Kaneko et al. (2024) show that Chain-of-Thought prompting reduces LLM gender bias

even on simple tasks. Popović and Lapshinova-Koltunski (2024) add another interesting result, showing that ChatGPT introduces different gender bias patterns when compared to MT systems. These findings suggest that model architecture plays a meaningful role in bias behavior, motivating cross-system comparative analyses.

2.2 Grammatical Gender in Chinese and Portuguese

Chinese has a semantic gender assignment system according to Corbett (1991), in which gender is only visible in the pronominal domain, and only in third-person pronouns in written Chinese. Consequently, there are many "personal dual gender" nouns (Quirk and Crystal, 2010), i.e., nouns whose gender is not encoded in the noun itself but must be inferred from context. Portuguese, by contrast, has a semantic-and-formal gender system (Corbett, 1991), in which gender is an inherent property of nouns and triggers agreement in various domains (Raposo et al., 2013). This asymmetry is a crucial motivation for our study, as it creates a systematic challenge for models performing translation between this language pair: the system must infer and assign grammatical gender using contextual cues that Chinese does not grammatically encode.

Chinese. Chinese is therefore typically classified as a language with covert gender. As an isolating language, Chinese words and characters do not present any morphological changes, independently of the head noun or any other constituents. This formal invariance leads to the classification of Chinese as a language with a pronominal gender system, i.e., a language where gender distinctions are manifested solely through personal pronouns (Corbett, 1991), and only in third-person pronouns in written text.

Chinese utilizes a tripartite distinction in its written third-person singular pronouns: 他 (*tā*, 'he') for masculine human referents, 她 (*tā*, 'she') for feminine human referents, and 它 (*tā*, 'it') for non-human entities, i.e., animals or inanimate objects. The corresponding plural forms are constructed simply by appending the plural morpheme '们 (*mén*)' to these singular bases. Despite these orthographic distinctions, the spoken form remains an identical homophone (*tā*). Thus, in oral communication Chinese does not offer any gender cues, relying entirely on context for referent identification.

Secondly, gender assignment in Chinese is predominantly determined by semantic criteria. Human referents are assigned gender according to their biological sex (male or female), while non-human referents fall into the neuter category. Human denoting nouns illustrate this semantic reliance in a particularly clear way. Lexical items such as 司机 ('driver'), 护士 ('nurse'), and 老师 ('teacher') are inherently underspecified with respect to gender. In language data involving this type of nouns, the noun itself remains invariant, and the gender of the referent is surfaced only through the gender of coreferential pronouns or explicit lexical modifiers.

When gender must be specified for clarity, Chinese employs lexical encoding rather than morphological inflection. For human referents, this is achieved by prefixing the noun with the morphemes 男 ('male') or 女 ('female'), as illustrated in (1):

(1) (a) 男司机 ('male driver') / 女司机 ('female driver')
    (b) 男护士 ('male nurse') / 女护士 ('female nurse')

Portuguese. The Portuguese language instantiates a pervasive grammatical gender system distinguishing masculine and feminine categories (Raposo et al., 2013). Gender is an inherent morphosyntactic property of the noun that governs agreement across the sentence.

Assignment follows both semantic and formal criteria (Kibort and Corbett, 2008): while nouns referring to male referents are typically masculine and those referring to female referents feminine, the gender of inanimate nouns is largely arbitrary (*copo* [masc., 'glass'] vs. *janela* [fem., 'window']), creating a significant challenge for computational models that must encode gender as a lexical property.

Occupation nouns exhibit several patterns. Some have different forms for masculine and feminine (*enfermeiro*/*enfermeira*, 'nurse'), others maintain a single form relying on the determiner to signal gender (*o*/*a estudante*, 'the student'). This variation is directly relevant to AT evaluation, as the surface form of the translated occupation noun constitutes the primary observable evidence of gender bias in translation output.

The true complexity of the Portuguese system lies in its agreement domain. Gender is not a

localized feature; it percolates through the entire nominal phrase and beyond. Determiners, including definite articles (*o*/*a*, 'the'), indefinite articles (*um*/*uma*, 'a'), and demonstratives (*este*/*esta*, 'this'; *aquele*/*aquela*, 'that'), must agree with the head noun in gender, as must adjectives with gendered forms. The numeral and pronominal systems are also partially gendered: cardinal numbers (*um*/*uma*, 'one'; *dois*/*duas*, 'two'), possessives (*meu*/*minha*, 'my'), and third-person pronouns (*ele*/*ela*, 'he/she') all carry gender distinctions. Gender agreement further extends to the verbal system: in passive constructions, past participles also agree in gender and number with the subject (*foi aberta* [fem.][sg.], 'was opened'). This level of syntactic integration means that a single gender assignment can dictate the morphological form of multiple subsequent words.

These structural properties create a fundamental challenge for Chinese–Portuguese AT: gender must be overtly expressed across multiple syntactic domains in the target, although it is entirely absent from the source. Failure to identify the correct gender value in the source produces agreement errors and may generate misinterpretation due to unintended bias in the translation output.

2.3 Linguistic Conditions as Gender Cues

Gender marks in languages are not fixed properties. At the same time, the gender assigned to a referent in the target language can shift depending on the linguistic information available in the source sentence.

The most extensively studied gender cue type is the coreferential pronoun. Stanovsky et al. (2019) built WinoMT on this principle: source sentences describe a scenario involving two occupation nouns, with a personal pronoun indicating the gender of one referent (e.g., *The developer argued with the designer because* ***she*** *did not like the design*).

These sentences are structurally ambiguous with respect to pronoun reference, requiring AT systems to resolve the referent without the pragmatic resources available to human readers. WinoMT thus constitutes a sensitive instrument for detecting gender bias: when a system defaults to a stereotypical gender assignment rather than following the pronominal cue, the bias becomes directly observable in the translation output. The interaction between pronoun cues and occupational stereotypes has since become a standard axis of evaluation in gender bias benchmarks (Levy et al., 2021; Sewunetie et al., 2024; Gkovedarou et al., 2025).

Beyond pronouns, sentence-level context has been shown to exert a meaningful influence on AT outputs. Hackenbuchner et al. (2024) compared gender translations of role names presented in isolation versus embedded in full sentences: out of context, 95% of role names were rendered in the generic masculine, whereas in sentence context this proportion dropped to 82%, with words such as *coordinator*, *mechanic*, and *musician* shifting gender depending on the surrounding context.

A third category of cue is the proper name. Although names function as implicit gender signals in many languages, they have been identified as unreliable carriers (Saunders and Olsen, 2023); they are therefore not considered as a condition in the present challenge set.

When the source is a language with a pronominal gender system such as Chinese, whether these cue types function in the same way remains an open empirical question that motivates the typology of linguistic conditions developed in Section 4.

3 Methodology

3.1 Translation Systems

To ensure broad coverage of the current automated translation landscape, eight systems were selected for evaluation: two commercial NMT engines and six LLMs. The two NMT systems (Google Translate and DeepL Translator) were accessed via API on February 26, 2026. The six LLMs were accessed via API on March 1 and 2, 2026, representing a diverse range of developer origins (United States, France, and China) and access modalities (proprietary and open-source): GPT-5.2 (OpenAI, hereafter GPT), Claude Sonnet 4.6 (Anthropic, hereafter Claude), Ministral-14B-2512 (Mistral AI, hereafter Ministral), Qwen3.5-397B-A17B (Alibaba, hereafter Qwen), Llama-3.3-70B-Instruct (Meta, hereafter Llama), and DeepSeek V3.2 (DeepSeek, hereafter DeepSeek). All systems were instructed to output European Portuguese.

For the LLM systems, translation was performed via API calls with temperature set to 0 across all models to minimize output variability and ensure reproducibility. Each sentence was submitted as an independent query with no prior con-

versational context, eliminating any potential influence of cross-sentence memory on gender output. All LLMs received an identical prompt: "*Translate the Chinese sentence into European Portuguese. Output only the translated text. Do not include quotes or explanations.*"

3.2 The Data

We used a purpose-built dataset (described in detail in Section 4) consisting of 495 Chinese sentences involving 45 occupational nouns (15 male-biased, 15 female-biased, and 15 unbiased) across 11 conditions encoding different gender cue types (Section 4.3). The 495 source sentences were translated into European Portuguese by the 8 systems described in Section 3.1, generating 3,960 target sentences.

3.3 Annotation Procedure

To evaluate the extent to which the eight models considered in this study were able to take into account different linguistic cues in the source sentence to achieve correct gender assignment in the target, the 3,960 outputs generated underwent an annotation process.

The annotation focused on the gender of all the elements in the translated noun phrase headed by the occupation target noun. Overall translation quality or syntactic correctness was not considered at this stage. This targeted approach reflects the focus of this study: determining which gender value does each AT system assign to the target noun, and whether this assignment aligns with the gender information present in the Chinese source.

Gender was determined firstly and primarily from the morphological form of the occupation noun in Portuguese. As all gendered elements in the noun phrase were considered, in cases where the noun form is invariant across genders, gender was inferred from the determiner (*o*/*a*, *os*/*as*, *este*/*esta*, *aquele*/*aquela*). Gender assignment was counted as correct if and only if the gender of all the elements in the target noun phrase matched the gender specified in the source, irrespective of number agreement or gender marking outside the nominal phrase.

Annotation was aided by a rule-based semi-automated procedure. A reference lexicon was compiled containing the masculine and feminine Portuguese surface forms of all 45 occupation nouns included in the challenge set (see Section 4). Each system output was automatically matched against this lexicon: tokens identified as masculine forms were labelled M, feminine forms were labelled F. Cases not recognized in the matching process, as well as occupation nouns with identical masculine and feminine Portuguese forms, were manually annotated by the first author, a trained linguist and native speaker of Chinese with advanced proficiency in Portuguese. All annotation labels were then reviewed by the same annotator to ensure consistency, and later validated by the second author, a native speaker of Portuguese.

4 Chinese Challenge Set Construction

Although a large-scale Chinese–Portuguese parallel corpus is available (Chao et al., 2018), it was compiled from news, legal, and general web sources and was not designed for gender-related linguistic research, rendering it unsuitable for gender bias evaluation without substantial curation. As no suitable dedicated challenge set exists for the Chinese–Portuguese language pair, we developed a methodology to generate such data from existing sources.

4.1 Data Source and Occupation Selection

To select the target nouns considered in this study, we drew on data from the U.S. Bureau of Labor Statistics (BLS) 2024 Current Population Survey (CPS), specifically on the information reported regarding: *Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity*[1], which provides nationally representative statistics on the gender composition of the U.S. labor force.

The BLS statistics are employed here as an operationalization criterion for classifying occupations into gender stereotype categories, as no comparable data is available for China or Portuguese-speaking countries. The classification serves to ensure a replicable and empirically grounded partition of occupations; it does not presuppose that the same distributions obtain in the cultural contexts most relevant to the source or target language of this study. The stereotype classification (see 4.2) was nonetheless validated against data available for Portugal[2] and the EU[3] which presents in-

[1] Available at https://www.bls.gov/cps/cpsaat11.htm. Accessed on February 22, 2026.

[2] Available at https://www.cig.gov.pt/wp-content/uploads/2023/11/BE2023trabalho.pdf. Accessed on May 15, 2026.

[3] Available at https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20180307-1. Accessed on May 15, 2026.

formation aggregated into groups of professional activities.

From the aforementioned data, we selected 45 occupations to achieve a balanced and representative distribution across the gender spectrum. Only occupations corresponding to lexicalized and widely recognized professional categories in Chinese and Portuguese were selected. The selection also prioritized occupations with a total employed population exceeding 100,000 workers, although three occupations with totals between 90,000 and 100,000 were included: 兽医 ('veterinarian', 98,000), 翻译 ('translator', 92,000), and 急救员 ('paramedic', 95,000), given their proximity to the threshold and their status as prototypical and widely recognized occupational categories.

4.2 Stereotyping Classification

To operationalize occupational gender stereotyping, we classified the 45 occupations selected into three groups based on the proportion of female workers (*female%*) as reported in the BLS dataset:

- Male-stereotyped (M): *female%* $< 40\%$ (e.g., 木匠 'carpenter', 4.2%; 总裁 'CEO', 33.0%)
- Neutral (N): $40\% \leq$ *female%* $\leq 60\%$ (e.g., 教授 'professor', 49.9%; 调酒师 'bartender', 54.3%)
- Female-stereotyped (F): *female%* $> 60\%$ (e.g., 护士 'nurse', 86.8%; 秘书 'secretary', 91.4%)

Each category contains 15 occupation nouns, yielding a balanced three-way partition of 15 male-stereotyped, 15 neutral, and 15 female-stereotyped occupation nouns. The thresholds of 40% and 60% were determined empirically to yield groups of equal size, ensuring comparability across experimental conditions. A narrower neutral band (e.g., 45–55%) would have produced an uneven distribution across categories, undermining the factorial balance of the design. The ± 10 percentage point margin around parity additionally avoids misclassifying near-parity occupations as stereotyped, a concern that arises when a strict 50% binary threshold is applied (Gkovedarou, 2025).

4.3 Sentence Design

To isolate the effect of occupational gender stereotyping from confounding lexical variables, all 495 sentences are constructed using a fixed set of predicate structures held constant across all 45 occupation nouns (see Table 1). Since the predicate is identical, any observed variation in translation outputs can be attributed to the occupation noun–gender cue interaction rather than to verb-specific gender associations. Each occupation noun is embedded in a matrix of 11 sentence types, described in the following paragraphs.

Base Condition. The base sentence (`base_null`) contains the occupation noun with no gendered pronouns or explicit gender markers, establishing a measure of the model's default gender assignment in the absence of any linguistic gender cue.

Explicit Gender Cue. In these two sentence types (`exp_m` and `exp_f`), the occupation noun is preceded by an explicit gender modifier (男/女, i.e., *male*/*female*), providing the model with maximal lexical evidence of the gender of the referent. These sentences test whether overt lexical gender marking is sufficient to override the model's stereotype prior—including in cases where the explicit gender modifier contradicts the gender stereotypically associated with the occupation noun (e.g., 男护士, *'male nurse'*, given that this profession is strongly female-stereotyped).

Coreference Conditions. The remaining eight sentence types are generated by orthogonally crossing three binary variables, yielding eight distinct conditions involving pronouns holding a coreference relation with the target noun. Each sentence type is assigned a semantic tag encoding all three dimensions simultaneously (e.g., `ante_cpx_f`):

The first factor is linear order. In sentences identified with `ante`, the occupation noun precedes the personal pronoun. In sentences identified with `post`, the personal pronoun precedes the occupation noun, requiring the model to resolve the reference of the pronoun retrospectively once it finds the occupation noun.

The second factor distinguishes simple and complex sentences. Simple sentences (`smp`) consist of a single clause with coreference being established within the same syntactic domain. Complex sentences (`cpx`) involve two clause structures that increase the syntactic distance between the occupation noun and the coreferential pronoun, placing

Tag	Chinese Sentence	English Gloss
base_null	我最近新认识了一名{occupation noun}。	I recently got acquainted with a(n) {occupation noun}.
exp_m	我最近新认识了一名男{occupation noun}。	I recently got acquainted with a **male** {occupation noun}.
exp_f	我最近新认识了一名女{occupation noun}。	I recently got acquainted with a **female** {occupation noun}.
ante_smp_m	这名{occupation noun}很满意他自己的工作环境。	This {occupation noun} is very satisfied with **his** work environment.
ante_smp_f	这名{occupation noun}很满意她自己的工作环境。	This {occupation noun} is very satisfied with **her** work environment.
ante_cpx_m	我最近新认识了一名{occupation noun}，他非常细心。	I recently got acquainted with a(n) {occupation noun}, and **he** is very meticulous.
ante_cpx_f	我最近新认识了一名{occupation noun}，她非常细心。	I recently got acquainted with a(n) {occupation noun}, and **she** is very meticulous.
post_smp_m	他是一名细心的{occupation noun}。	**He** is a meticulous {occupation noun}.
post_smp_f	她是一名细心的{occupation noun}。	**She** is a meticulous {occupation noun}.
post_cpx_m	因为他表现得非常细心，所以大家才如此信任这名{occupation noun}。	Because **he** demonstrates such meticulousness, everyone trusts this {occupation noun} so much.
post_cpx_f	因为她表现得非常细心，所以大家才如此信任这名{occupation noun}。	Because **she** demonstrates such meticulousness, everyone trusts this {occupation noun} so much.

Table 1: The eleven sentence templates used in the challenge set. Gender-bearing elements (modifiers and pronouns) are shown in **bold** (Latin script) and underlined (Chinese script). {occupation noun} denotes the placeholder replaced by each of the 45 occupation nouns.

greater demand on the coreference resolution capacity of the model.

The third factor identifies the gender of the personal pronoun as either masculine (m) (他, *'he'*) or feminine (f) (她, *'she'*).

In total, there are 11 sentences for each occupation noun following the aforementioned templates, yielding a corpus of $45 \times 11 = 495$ source sentences. The complete sentence matrix is illustrated in Table 1. All 495 Chinese source sentences were reviewed for naturalness and grammaticality by the first and third authors, both native speakers of Chinese.

5 Results

All results are reported as gender accuracy percentages across 45 sentences per condition per system (N = 45), with masculine (M) and feminine (F) outputs evaluated separately against the gender value expected given each source sentence condition. Two 2gen outputs were identified in the dataset: instances in which a system produced a form with ambiguous or inclusive gender marking, for example, *um(a) tripulante* ('a crew member'), *este/a empregado/a* ('this employee'). These cases are reported in footnotes where they occur.

Tables 2, 3, and 4 report gender output distributions, masculine–feminine asymmetry, and inferential statistics, respectively.

5.1 Base Condition

Under the base_null condition, source sentences did not contain gender information of any kind. Therefore, gender assignment in the Portuguese output reflects each system's default gender associated with each noun in our dataset.

All eight systems exhibited a strong masculine default bias (Table 2). Masculine outputs ranged from 77.8% (Ministral) to 93.3% (DeepSeek), with a cross-system mean of 86.1%. Feminine outputs were consistently marginal, never exceeding 22.2% for any system. This experimental condition establishes the base results against which the effect of explicit and implicit gender cues in subsequent conditions is assessed.

5.2 Explicit Gender Cues

When prenominal gender modifiers 男/女 ('male/female') immediately preceding the occupation noun were used, all eight systems achieved 100.0% accuracy on both target genders, including anti-stereotypical cases (e.g., 女工程师, *'female engineer'*), indicating that overt lexical gender

Model	base_null	ante_smp		ante_cpx		post_smp		post_cpx	
	M%	M Acc.%	F Acc.%	M Acc.%	F Acc.%	M Acc.%	F Acc.%	M Acc.%	F Acc.%
Google Translate	82.2	100.0	100.0	77.8	28.9	100.0	100.0	100.0	80.0
DeepL Translator	86.7	100.0	100.0	88.9	24.4	100.0	100.0	100.0	97.8
GPT-5.2	82.2[†]	100.0	100.0	95.6	100.0	100.0	100.0	97.8	100.0
Claude Sonnet 4.6	88.9	100.0	100.0	97.8	100.0	100.0	100.0	100.0	100.0
Ministral-14B	77.8	82.2[‡]	82.2	80.0	91.1	100.0	100.0	91.1	48.9
Qwen3.5-397B	91.1	100.0	100.0	97.8	100.0	100.0	100.0	100.0	100.0
Llama3.3-70B	86.7	100.0	88.9	86.7	62.2	100.0	100.0	97.8	28.9
DeepSeek V3.2	93.3	93.3	64.4	97.8	97.8	100.0	100.0	93.3	95.6
Mean	**86.1**	**97.0**	**92.0**	**90.3**	**75.5**	**100.0**	**100.0**	**97.5**	**81.4**

Table 2: Gender output distribution under the base (base_null) and all other conditions, by system (N=45 per condition per system). M% (base_null) = proportion of masculine outputs in sentences without any gender information. M/F Acc.% = proportion of outputs in which the occupation noun phrases are correctly rendered in masculine/feminine form under the respective linguistic conditions. All explicit gender cue conditions (exp_m and exp_f) yielded 100.0% accuracy across all eight systems and are, thus, omitted from the table.
[†]GPT produced one 2gen output under base_null, counted as a non-masculine response. [‡]Ministral produced one 2gen output under ante_smp_m, counted as incorrect for M Acc.%.

expression constitutes a fully sufficient cue that no system overrides based on the gender-stereotype bias of the model.

5.3 Coreference Conditions

Results pertaining to gender information rendered by coreference relations are examined along each of the three factors tested in our data: pronoun position (occurring before or after the occupation noun), syntactic complexity (simple and complex sentences), and gender of the target noun (masculine and feminine).

Position. Conditions in which the gendered pronoun precedes the target occupation noun (post) consistently outperform their counterparts in which the target noun precedes the pronoun (ante) across both complexity levels: under simple sentences, post achieves 100.0% for both M and F, while ante reaches 97.0% (M) and 92.0% (F); under complex sentences, post also shows higher accuracy than ante for both genders (M: 97.5% vs. 90.3%; F: 81.4% vs. 75.5%). This advantage is consistent with left-to-right token processing: when the gender-carrying pronoun occurs before the gender unmarked noun in the source sentence, it is encoded into the contextual representation before the occupation noun is reached, providing a stronger prior for gender selection and reducing the window in which occupational stereotype defaults can take effect.

Complexity. Syntactic complexity produces the most pronounced performance drop in the conditions included in the dataset. In simple sentence structures, six of the eight models achieve perfect or near-perfect performance regardless of target gender or pronoun position. However, in complex sentence structures, performance degrades substantially, with the greatest variation observed when a feminine pronoun occurs after the target noun (ante_cpx_f): Google Translate falls to 28.9% and DeepL to 24.4%, meaning that in more than 70% of cases these systems fail to render the feminine gender information conveyed by the pronoun in the source sentence. By contrast, GPT, Claude, and Qwen maintain accuracy at or above 97.8% across all complex conditions, demonstrating robust coreference resolution under increased syntactic distance and with lower sensitivity to linear order.

Target gender. Under simple conditions, the gap between masculine and feminine accuracy is modest: 5.0 percentage points in ante conditions (occupation noun precedes pronoun) and no difference in post conditions (pronoun precedes occupation noun). Under complex conditions, this gap widens notably: 14.8 and 16.1 percentage points for ante and post, respectively. In other words, complex sentence structures do not merely reduce accuracy overall; the reduction is greater for feminine cues than for masculine cues.

This pattern suggests that gender asymmetry is a latent tendency among the eight models tested, that is amplified when syntactic complexity increases: under simple conditions, systems handle both genders adequately; under complex conditions, they retreat towards masculine default.

Model	Δ_{aSmp}	Δ_{aCpx}	Δ_{pSmp}	Δ_{pCpx}	$\overline{\Delta}$
Google Translate	+0.0	+48.9	+0.0	+20.0	+17.2
DeepL Translator	+0.0	+64.5	+0.0	+2.2	+16.7
GPT-5.2	+0.0	−4.4	+0.0	−2.2	−1.7
Claude Sonnet 4.6	+0.0	−2.2	+0.0	+0.0	−0.6
Ministral-14B	+0.0	−11.1	+0.0	+42.2	+7.8
Qwen3.5-397B	+0.0	−2.2	+0.0	+0.0	−0.6
Llama3.3-70B	+11.1	+24.5	+0.0	+68.9	+26.1
DeepSeek V3.2	+28.9	+0.0	+0.0	−2.3	+6.7
Mean	+5.0	+14.8	+0.0	+16.1	+9.0

Table 3: Masculine–feminine accuracy asymmetry (Δ = M Acc.% − F Acc.%, in percentage points) by system and coreference condition. Positive values indicate superior masculine accuracy; negative values indicate superior feminine accuracy. Δ_{aSmp} = `ante_smp`; Δ_{aCpx} = `ante_cpx`; Δ_{pSmp} = `post_smp`; Δ_{pCpx} = `post_cpx`; $\overline{\Delta}$ = mean across the four conditions.

Predictor	β	OR	95% CI	
Ref.: M gender, `smp`, `ante`, NMT, neutral				
Target gender: F	−1.421	0.24	[0.17, 0.33]	***
Complexity: `cpx`	−1.840	0.16	[0.11, 0.23]	***
Position: `post`	+0.927	2.53	[1.87, 3.41]	***
System type: LLM	+0.775	2.17	[1.61, 2.92]	***
Stereotype: M	−0.652	0.52	[0.35, 0.78]	**
Stereotype: F	−0.245	0.78	[0.51, 1.19]	
N=2,880 AIC=1376.9 σ^2_{occ}=0.0758				

Table 4: Mixed-effects logistic regression predicting correct gender accuracy in coreference conditions (N=2,880; occupation included as random intercept). β = log-odds coefficient; OR = odds ratio; 95% CI = confidence interval for OR. Reference category: masculine target gender, `smp`, `ante`, NMT system, gender-neutral stereotype. ** $p < .01$; *** $p < .001$.

5.4 M–F Asymmetry Across Conditions

Table 3 quantifies the masculine–feminine accuracy gap per system and condition, complementing the aggregate patterns discussed above with system-level detail. The cross-system mean confirms the complexity-driven amplification identified in Section 5.3: mean Δ remains substantially lower under simple conditions (Δ_{pSmp}=+0.0, Δ_{aSmp}=+5.0) than under complex ones (Δ_{aCpx}=+14.8, Δ_{pCpx}=+16.1).

At the system level, Google Translate (Δ_{aCpx}=+48.9) and DeepL (Δ_{aCpx}=+64.5) show the largest masculine advantage under `ante_cpx` condition, while Llama records the highest overall asymmetry ($\overline{\Delta}$=+26.1). By contrast, GPT, Claude, and Qwen show marginally negative $\overline{\Delta}$ values, indicating effectively symmetric gender cue processing across conditions.

5.5 Predictors of Gender Accuracy

To assess whether the observed patterns reflect statistically stable effects, we fitted a mixed-effects logistic regression model predicting gender accuracy across all coreference conditions (base condition excluded as gender-indeterminate; explicit conditions excluded due to 100% accuracy with zero variance; N=2,880). Five predictors were entered as fixed effects: target gender, syntactic complexity, occupation noun position, system type, and occupational gender stereotype, with occupation as a random intercept to account for dependence among observations sharing the same occupation noun across conditions. Results are reported as odds ratios (OR) with 95% confidence intervals (Table 4).

All predictors except F-stereotyped occupation reached statistical significance: target gender, syntactic complexity, occupation noun position, and system type at $p < .001$, and M-stereotyped occupation at $p < .01$. Syntactic complexity produces the largest effect (OR=0.16, CI [0.11, 0.23]): complex sentences reduce the odds of accuracy to just 16% of those under simple conditions, confirming it as the strongest individual predictor of translation failure.

Target gender is the second strongest predictor (OR=0.24, CI [0.17, 0.33]): feminine-cued sentences are statistically harder to resolve correctly than masculine-cued ones across all models and conditions, providing inferential confirmation of the asymmetry documented in the previous section.

Conditions in which the pronoun precedes the occupation noun (`post`) show higher accuracy than the reverse order (`ante`) (OR=2.53, CI [1.87, 3.41]). And LLMs outperform NMT systems (OR=2.17, CI [1.61, 2.92]).

Among stereotype categories, M-stereotyped occupations show significantly lower accuracy than neutral ones (OR=0.52, CI [0.35, 0.78]), indicating that masculine occupational stereotypes influence gender assignment even when the source contains a gendered pronoun; F-stereotyped occupations do not differ significantly from neutral ones ($p = .253$). Finally, a supplementary test of the gender × complexity interaction did not reach significance (χ^2=1.508, $p = .219$), confirming that the two factors contribute independently to translation difficulty without compounding each other.

6 Final Remarks

This study introduced a Chinese challenge set for evaluating gender bias in Chinese–Portuguese automated translation, comprising 495 source sen-

tences generated from 45 occupation nouns across 11 sentence templates and evaluated across two NMT systems and six LLMs. The results address the three objectives guiding this work.

Regarding the second goal of this research, the data reveal a hierarchy of gender cue effectiveness in Chinese source sentences. Explicit prenominal gender modifiers (男/女, 'male/female') constitute the most reliable cue type, achieving 100.0% accuracy across all systems. Simple coreferential pronouns (他/她, 'he/she') are generally sufficient for most systems, though three systems showed localized failures under anaphoric conditions. Complex sentence conditions proved the most demanding, with feminine cue accuracy falling as low as 24.4% for DeepL (`ante_cpx_f`).

In this way, different linguistic conditions produce systematically different gender bias results across systems, with target gender, syntactic complexity, and pronoun position each emerging as significant independent predictors of correct gender accuracy in the mixed-effects model ($p < .001$).

Regarding our third objective, the proposed methodology produces a valid evaluation instrument: it successfully differentiates both between systems and between conditions, with masculine default rates under gender-indeterminate conditions (77.8–93.3%) consistent with prior literature (Stanovsky et al., 2019; Prates et al., 2020), and complex sentence conditions producing the maximal gap between systems observed in the study.

A notable finding concerns the divergence between NMT systems and LLMs. Google Translate and DeepL exhibited pronounced masculine bias under complex conditions, defaulting to masculine output even when an explicit feminine pronoun was present in the source. By contrast, GPT, Claude, and Qwen maintained near-symmetric accuracy across all coreference conditions, suggesting that these systems are better equipped to resolve long-distance coreference. Whether this advantage is attributable to model scale, instruction tuning, or training data composition remains an open question.

Limitations and Future Work. This study operates within a binary gender framework reflecting standard European Portuguese morphology; gender-inclusive strategies fall outside the scope of this paper. Despite the use of available data for Portuguese, occupational stereotype classifications derive from U.S. statistics, and may not reflect the gender–occupation associations in Chinese and Portuguese-speaking contexts. Annotation is restricted to the noun phrase in which the target noun occurs and does not capture broader agreement or overall translation quality, including target-variety compliance.

Future work should extend gender annotation to sentence-level agreement and evaluate overall translation quality. We will also expand the occupation list and sentence templates, incorporate naturally occurring sentences to complement the controlled template-based design, and develop finer-grained measures of syntactic complexity.

Bias Statement. This paper addresses representational harm arising from gender stereotyping in Chinese–Portuguese AT. When AT systems default to masculine forms for gender-indeterminate occupation nouns or fail to transmit gender cues in the source, they introduce gendered interpretations that were absent from the source, potentially reinforcing gender stereotypes and rendering women and non-binary individuals less visible in translated professional discourse. This harm is compounded in the Chinese–Portuguese direction specifically, where masculine defaults introduce bias that was absent from the gender-neutral source. We acknowledge that operating within a binary gender framework is itself a normative limitation; extending gender bias evaluation beyond the masculine/feminine binary remains an important direction for future work.

Acknowledgements

We acknowledge the support of Fundação para a Ciência e a Tecnologia (UID/00214: Centro de Linguística da Universidade de Lisboa).

Carbon Impact Statement

This study did not involve training any machine learning models. All automated translation outputs were obtained via API calls to eight externally hosted systems (Google Translate, DeepL, and six LLMs accessed through OpenRouter), comprising a total of 3,960 inference queries. As the computational infrastructure is operated by third-party providers, precise energy consumption figures are not available to the authors. The overall computational footprint of this study is estimated to be minimal relative to model training workloads.

References

Blodgett, Su Lin, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5454–5476.

Bonifácio, Sofia, Helena Moniz, and Luisa Coheur. 2025. Diving into gender translation bias for the portuguese language. In *28th European Conference on Artificial Intelligence, ECAI 2025, including 14th Conference on Prestigious Applications of Intelligent Systems, PAIS 2025*, pages 1067–1074. IOS Press BV.

Chao, Lidia S, Derek F Wong, Chi Hong Ao, and Ana Luısa Leal. 2018. Um-pcorpus: a large portuguese-chinese parallel corpus. In *Proceedings of the LREC 2018 Workshop "Belt & Road: Language Resources and Evaluation*, pages 38–43.

Cho, Won Ik, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181.

Corbett, Greville G. 1991. *Gender*. Cambridge University Press.

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*, volume 36, pages 11855–11863.

Fleisig, Eve and Christiane Fellbaum. 2022. Mitigating gender bias in machine translation through adversarial learning. *arXiv preprint arXiv:2203.10675*.

Font, Joel Escudé and Marta R Costa-Jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154.

Gkovedarou, Eleni, Joke Daems, and Luna De Bruyne. 2025. Gender bias in English-to-Greek machine translation. In Hackenbuchner, Janiça, Luisa Bentivogli, Joke Daems, Chiara Manna, Beatrice Savoldi, and Eva Vanmassenhove, editors, *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 17–45, Geneva, Switzerland, June. European Association for Machine Translation.

Hackenbuchner, Janiça, Joke Daems, Arda Tezcan, and Aaron Maladry. 2024. You shall know a word's gender by the company it keeps: Comparing the role of context in human gender assumptions with MT. In Savoldi, Beatrice, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 31–41, Sheffield, United Kingdom, June. European Association for Machine Translation (EAMT).

Kaneko, Masahiro, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

Kibort, Anna and Greville G Corbett. 2008. Grammatical features inventory. *Typology of grammatical features*.

Levy, Shahar, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 2470–2480.

Piergentili, Andrea, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. Hi guys or hi folks? benchmarking gender-neutral machine translation with the gente corpus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140.

Popović, Maja and Ekaterina Lapshinova-Koltunski. 2024. Gender and bias in amazon review translations: by humans, mt systems and chatgpt. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30.

Prates, Marcelo OR, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381.

Quirk, Randolph and David Crystal. 2010. *A comprehensive grammar of the English language*. Pearson Education India.

Raposo, Eduardo Buzaglo, Maria Fernanda Nascimento, Maria Antónia Mota, Luísa Segura, and Amália Mendes. 2013. *Gramática do Português*. Fundação Calouste Gulbenkian, Lisboa.

Sant, Aleix, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in mt with llms. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139.

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

Saunders, Danielle and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 85–93.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Savoldi, Beatrice, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076.

Sewunetie, Walelign, Atnafu Tonja, Tadesse Belay, Hellina Hailu Nigatu, Gashaw Gebremeskel, Zewdie Mossie, Hussien Seid, and Seid Yimam. 2024. Gender bias evaluation in machine translation for Amharic, Tigrigna, and afaan oromoo. In Savoldi, Beatrice, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 1–11, Sheffield, United Kingdom, June. European Association for Machine Translation (EAMT).

Sharma, Shanya, Manan Dey, and Koustuv Sinha. 2022. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1968–1984.

Stanovsky, Gabriel, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1679–1684.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2979–2989.

Appendix A. List of occupation nouns in the dataset.

Table 5 lists the 45 occupation nouns included in the challenge set, ordered by female percentage as reported in the U.S. Bureau of Labor Statistics.

#	Chinese	English	Total (thousands)	Female%	Label
1	木匠	Carpenters	1,279	4.2	M
2	机械工	Machinists	281	4.8	M
3	消防员	Firefighters	298	6.2	M
4	飞行员	Aircraft pilots and flight engineers	213	9.6	M
5	出租车司机	Taxi drivers	514	12.6	M
6	警察	Police officers	678	14.2	M
7	程序员	Computer programmers	334	17.8	M
8	主厨	Chefs and head cooks	515	21.6	M
9	急救员	Paramedics	95	25.3	M
10	洗碗工	Dishwashers	277	26.9	M
11	建筑师	Architects	209	27.4	M
12	总裁	Chief executives	1,728	33.0	M
13	公交车司机	Bus drivers, transit and intercity	271	35.8	M
14	清洁工	Janitors and building cleaners	2,304	35.9	M
15	牙医	Dentists	169	38.8	M
16	律师	Lawyers	1,146	42.0	N
17	厨师	Cooks	2,005	42.5	N
18	摄影师	Photographers	210	43.5	N
19	销售员	Retail salespersons	2,734	47.5	N
20	教练	Coaches and scouts	340	48.5	N
21	教授	Postsecondary teachers	1,072	49.9	N
22	艺术家	Artists	318	53.8	N
23	医学研究员	Medical scientists	127	54.0	N
24	调酒师	Bartenders	450	54.3	N
25	生物学家	Biological scientists	105	54.7	N
26	电子装配工	Electrical and electronics assemblers	115	54.9	N
27	房地产经纪人	Real estate brokers and sales agents	1,016	56.7	N
28	会计	Accountants	1,778	57.2	N
29	审计	Auditors	1,778	57.2	N
30	药剂师	Pharmacists	326	59.7	N
31	作家	Writers and authors	257	64.7	F
32	面包师	Bakers	243	67.1	F
33	服务员	Waiters and waitresses	1,786	67.8	F
34	兽医	Veterinarians	98	68.6	F
35	收银员	Cashiers	2,569	68.9	F
36	编辑	Editors	132	72.6	F
37	心理医生	Psychologists	215	74.0	F
38	翻译	Interpreters and translators	92	74.8	F
39	空乘	Flight attendants	129	77.2	F
40	老师	Teachers	3,632	77.7	F
41	护士	Registered nurses	3,571	86.8	F
42	家政	Maids and housekeeping cleaners	1,413	87.7	F
43	助理	Assistants	1,758	91.4	F
44	秘书	Secretaries and administrative assistants	1,758	91.4	F
45	营养师	Dietitians and nutritionists	142	92.1	F

Table 5: Occupation nouns included in the Chinese Challenge Set, ordered by the percentage of female professionals (BLS 2024 Current Population Survey). Total employment figures are in thousands. M = female% $< 40\%$; N = $40\% \leq$ female% $\leq 60\%$; F = female% $> 60\%$. Each stereotype category contains exactly 15 occupation nouns.