FairTranslate: An English-French Dataset for Gender Bias Evaluation in Machine Translation by Overcoming Gender Binarity

Fanny Jourdan, Yannick Chevalier, Cécile Favre

FairTranslate: An English-French Dataset for Gender Bias Evaluation in Machine Translation by Overcoming Gender Binarity

Fanny Jourdan
fanny.jourdan@irt-saintexupery.com
IRT Saint Exupery
Toulouse, France

Yannick Chevalier
yannick.chevalier@univ-lyon2.fr
Université Lumière Lyon 2
IHRIM UMR 5317
Lyon, France

Cécile Favre
cecile.favre@univ-lyon2.fr
Université Lumière Lyon 2, Université Claude Bernard Lyon 1, ERIC
69007, Lyon, France

Abstract

Large Language Models (LLMs) are increasingly leveraged for translation tasks but often fall short when translating inclusive language – such as texts containing the singular '*they*' pronoun or otherwise reflecting fair linguistic protocols. Because these challenges span both computational and societal domains, it is imperative to critically evaluate how well LLMs handle inclusive translation with a well-founded framework.

This paper presents FairTranslate, a novel, fully human-annotated dataset designed to evaluate non-binary gender biases in machine translation systems from English to French. FairTranslate includes 2418 English-French sentence pairs related to occupations, annotated with rich metadata such as the stereotypical alignment of the occupation, grammatical gender indicator ambiguity, and the ground-truth gender label (male, female, or inclusive).

We evaluate four leading LLMs (Gemma2-2B, Mistral-7B, Llama3.1-8B, Llama3.3-70B) on this dataset under different prompting procedures. Our results reveal substantial biases in gender representation across LLMs, highlighting persistent challenges in achieving equitable outcomes in machine translation. These findings underscore the need for focused strategies and interventions aimed at ensuring fair and inclusive language usage in LLM-based translation systems.

We make the FairTranslate dataset publicly available on **Hugging Face**, and disclose the code for all experiments on **GitHub**.

CCS Concepts

• **Computing methodologies** → **Machine translation**.

Keywords

Fairness, Natural Language Processing, Translation, LLM, Gender

ACM Reference Format:
Fanny Jourdan, Yannick Chevalier, and Cécile Favre. 2025. FairTranslate: An English-French Dataset for Gender Bias Evaluation in Machine Translation by Overcoming Gender Binarity. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3715275.3732013

FAccT '25, June 23–26, 2025, Athens, Greece

ACM ISBN 979-8-4007-1482-5/2025/06
https://doi.org/10.1145/3715275.3732013

1 Introduction

The advent of Large Language Models (LLMs) has significantly advanced machine translation, enabling cross-linguistic communication at an unprecedented scale. However, these systems often exhibit and perpetuate social biases, particularly when handling gendered language [19]. While much of the prior work has focused on evaluating binary gender biases, where the primary challenge is determining whether a source sentence refers to masculine or feminine forms, recent developments in inclusive linguistic practices have introduced an additional layer of complexity: accurate translation of gender-inclusive expressions.

A key translation challenge lies in the handling of the English singular '*they*', which serves as an inclusive non-binary or unspecified gender pronoun. Unlike traditional binary gendered pronouns, the singular '*they*' requires models to both infer that the pronoun is singular, and to preserve this gender-neutrality in their translations. Furthermore, in traditionally gendered languages such as French, articles and noun endings (for occupations, for example) are also often gendered and hence pose a challenge for fair machine translation. In addition to these two challenges, LLM translation should be able to generate gender-neutral neopronouns such as the French '*iel*' (combining gender-specific '*il*' (he) and '*elle*' (she)). Figure 1 illustrates these challenges by comparing translation errors in binary and non-binary gender scenarios, highlighting the additional complexity introduced by inclusive language.

In this work, we propose a structured approach to evaluate gender biases in machine translation. We develop and publicly share FairTranslate, a novel English-French translation dataset designed for assessing both binary and non-binary gender biases. The dataset is composed of sentences related to specific occupations. It is richly annotated with metadata, including gender labels (male, female, or inclusive), markers of gender ambiguity (in the english sentence), and indicators of alignment with stereotypical occupational gender-roles. This detailed annotation enables fine-grained analyses of model performance, particularly in their treatment of inclusive language.

Using FairTranslate, we conduct an evaluation of four widely used LLMs (Gemma2 [29], Mistral [18], Llama3.1, and Llama3.3 [11]) across four prompting strategies, including moral and linguistic promptings. Our findings reveal that current machine translation

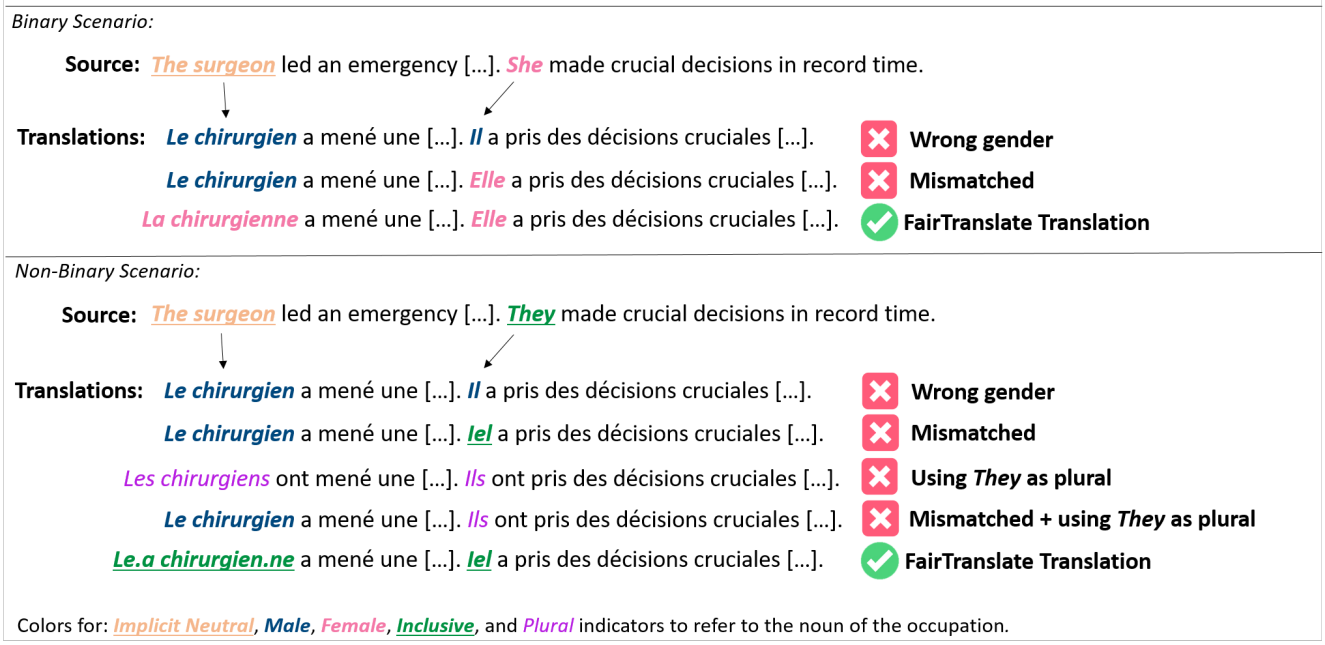

Figure 1: Illustration of potential translation errors in handling gendered language in machine translation. Starting from a source sentence in English, the binary scenario includes errors of incorrect gender assignment and mixed-gender outputs (combining correct and incorrect gender). The non-binary scenario extends these errors to include treating the singular '*they*' as plural, and mixing an incorrect gender assignement with a plural. The final line shows the reference translation from the FairTranslate dataset, exemplifying an accurate treatment of inclusive language.

systems struggle significantly with inclusive language. Translation quality for inclusive forms consistently lags behind that of binary forms. We observe that French inclusive markers such as '*iel*' are almost never generated, even with tailored prompting. Importantly, we show that these failures are not solely due to the recent introduction of inclusive practices into the French language. Instead, they reflect deeper issues in the ability of LLMs to represent, interpret and translate established inclusive constructs in English, such as the singular '*they*'.

2 Background

From a French linguistic point of view, grammatical gender is a tool that operates according to two main principles [8]:

Principle A: Grammatical gender ensures the internal cohesion of sentences. A donor element (the noun, which has a fixed gender) imposes gender agreement on satellite elements (adjectives, determiners, pronouns) [6].

Principle B: In the case of nouns referring to human beings (such as in occupation titles), grammatical gender helps convey the gender identity of the person who is being discussed.

Depending on the language, these two main principles may be applied differently. This is particularly evident in the comparison between French (a language with highly pronounced grammatical gender) and English (a language where grammatical gender is considered less prominent) [6]. The challenges of developing non-discriminatory writing protocols are therefore different in the two languages. While the non-discriminatory protocols of English are relatively stable and grammaticalized, they are less established in French where grammatical gender is more complex [1, 24]. The challenges which are introduced by such differences in grammatical gender across languages are significant and must be explicitly addressed. Specifically:

(i) In English, *principle A* does not apply. The gender of nouns referring to human beings does not affect the satellite elements that refer to the noun (e.g., determiners: '*a nurse, the nurse*'; adjectives: '*the professional nurse*'; or verb forms: '*the nurse has come*'). In contrast, in French, the grammatical gender of the noun influences satellite elements such as determiners (*'un'/'une', 'le'/'la'*), adjectives (*'professionnel'/'professionnelle'*), and compound verb forms (*'il est venu'/'elle est venue'*) [12].

(ii) In English, *principle B* rarely applies as most occupations names tend to be gender-invariant. In fact, the gender of nouns referring to human beings is rarely marked in English (e.g., '*nurse*'). However, in French, gender of nouns referring to human being is often marked through alternating suffixes (e.g., *'infirmier'* meaning male nurse, and *'infirmière'* meaning female nurse) [2].

(iii) Gender in pronouns is marked differently in English and French. In English, third-person singular pronouns (*'he'/'she'*) indicate gender, but the third-person plural pronoun (*'they'*) does not. In contrast, French marks gender in both third-person singular (*'il'/'elle'*) and plural (*'ils'/'elles'*) forms. Additionally, English possessive determiners inherit the gender of their antecedent (e.g., '*his/her dog*' from '*he/she owns a dog*'), whereas French possessives agree with the gender of the noun they determine (e.g., '*son chien*' for a male dog and '*sa chienne*' for a female dog, regardless of the owner's gender).

(iv) Finally, there is a particular difficulty in English in distinguishing between the two main uses of *they*: first as a third-person plural pronoun (*Plural They*) and second as an inclusive third-person singular pronoun (*Singular They* or *Inclusive They*). This issue does not arise in French, where separate pronouns are available for third-person singular ('*iel*' or equivalents) and third-person plural ('*iels*' or equivalents) to avoid expressing grammatical gender.

Machine translation from English to French must therefore propose satisfactory solutions to the following challenges:

(i') Once the correct French target word has been identified, the machine translation system must apply gender agreement rules to all related satellite elements (determiners, adjectives, verb forms), as well as to the pronouns used.

(ii') For English source terms (e.g., *'nurse'*), machine translation systems should generate three corresponding forms in French: the masculine form (*'infirmier'*), the feminine form (*'infirmière'*), and the inclusive form (*'infirmier.ière'* or an equivalent). This selection can be guided by contextual cues in the English text, such as pronouns (*'he'/'she'/inclusive 'they'*) and possessive determiners (*'his'/'her'/inclusive 'their'*). This process is analogous to performing coreference resolution: the task of identifying all expressions in a text that refer to the same entity, often focusing on how pronouns are linked to their antecedents. However, even when such information is available—and especially when no contextual cues are present to guide the model—the system may fail to appropriately gender the target term. This issue partly stems from the historical perception of the masculine form as "neutral" in the French linguistic tradition, aligned with the conservative recommendations of the *Académie Française*. Nonetheless, the widespread claim that feminine forms do not always exist is inaccurate: the morphology of French is highly flexible in forming feminine forms, as evidenced by their presence in the earliest written texts in the language[1]. The tendency of models to predominantly use masculine forms (or feminine forms for certain highly stereotyped professions) reinforces existing gendered representations [15]. It is crucial to prevent quantitative descriptions—which are themselves products of social constructions—from acting as descriptive norms that evolve into prescriptive norms. This phenomenon is particularly significant in contexts such as influencing young people's career choices, where such biases may perpetuate gender stereotypes [16].

(iii') and (iv') The differences in pronoun usage between English and French require machine translation systems to distinguish between singular '*they*' (person 3) and plural '*they*' (person 6). This distinction involves accurately identifying whether the antecedent in the English source text is singular or plural. Once this has been determined, translating into French is generally straightforward for singular '*they*' (*Inclusive They* = '*iel*'). However, difficulties persist for plural 'they,' as the system must select among the three available forms in French (*'ils'/'elles'/'iels'*) based on the desired level of inclusivity.

3 Related Work

Evaluating gender bias in Machine Translation (MT) has been extensively studied using benchmarks like Winogender [27], Winobias [33], WinoMT [28] and, more recently, WinoPron [14], which assess coreference resolution to determine how models attribute gender based on context. Similarly, MT-GenEval [9] employs a counterfactual methodology by systematically modifying sentences to analyze gender biases. Our approach integrates both coreference resolution and counterfactual evaluation, extending these methodologies beyond the binary framework to include non-binary gender and address the broader challenges associated with it.

Recent studies addressing non-binary gender [10, 13, 31] focus on evaluating bias related to gender and LGBTQ+ identities, but do not specifically target machine translation. AmbGIMT [5] evaluates non-binary gender in English-Chinese MT, but its focus is on attitude translation rather than coreference or the use of singular they. An English-German MT study also exists, focusing on gender-neutral person-referring terms [20].

While these works cover English-Chinese and English-German, the addition of an English-French dataset enables exploration of non-binary gender translation across typologically diverse languages such as French, German, and Chinese, opening up new research opportunities.

4 FairTranslate Dataset

Building on the considerations outlined in earlier sections, the FairTranslate Dataset includes **2,418 entries** designed to investigate how LLMs handle gender in English-to-French translation, paying special attention to inclusive forms and subtler expressions of gender bias.

FairTranslate is composed of English-French sentence pairs, each centered around an occupation to investigate how gender is expressed and translated. Each sentence is annotated with a gender label (male, female, or inclusive) corresponding to the individual referenced by the occupation. In addition, the dataset includes rich metadata, detailed in the following section.

To enable counterfactual interventions and direct comparisons, each example appears in all three gender variants (male, female, and inclusive). This design facilitates the evaluation of gender-specific translations and supports research on coreference resolution. Examples of sentences from the dataset, along with their corresponding annotations, are illustrated in Figure 2.

In combination, these design choices produce a dataset that can serve as a robust benchmark for evaluating fairness, inclusivity, and nuanced coreference resolution in modern translation models. The result is a richer, more controlled environment for studying fairness and inclusivity in machine translation.

4.1 Structure

Each entry in the FairTranslate dataset corresponds to a pair of English-French sentence translations, linked to an occupation, and annotated with several attributes. These attributes reflect the key variables in our study, enabling counterfactual comparisons and nuanced analyses of model behavior.

- **English**: English sentences which involve an occupation. These sentences were chosen to examine how models respond to both entrenched stereotypes as well as to more nuanced contexts.

[1] See a list of all historically attested feminine forms in French at https://siefar.org/la-guerre-des-mots/presentation/

English	French	Gender	Ambiguity	Stereotype	Occupation
This dietitian is *known* for [...] needs of ***his*** patients	***Ce diététicien*** est ***reconnu*** pour [...] de *ses* patients.	male	unambiguous	female-stereotyped	['***diététicien***','*diététicienne*', '***diététicien.ne***']
This dietitian is *known* for [...] needs of ***her*** patients	*Cette diététicienne* est ***reconnue*** pour [...] de *ses* patients.	female	unambiguous	female-stereotyped	['***diététicien***','*diététicienne*', '***diététicien.ne***']
This dietitian is *known* for [...] needs of ***their*** patients	***Ce.tte diététicien.ne*** est ***reconnu.e*** pour [...] de *ses* patients.	inclusive	unambiguous	female-stereotyped	['***diététicien***','*diététicienne*', '***diététicien.ne***']
An engineer works at an industrial [...] new equipment.	***Un ingénieur*** travaille sur un site industriel [...] équipements.	male	ambiguous	male-stereotyped	['***ingénieur***','*ingénieure*', '***ingénieur.e***']
An engineer works at an industrial [...] new equipment.	*Une ingénieure* travaille sur un site industriel [...] équipements.	female	ambiguous	male-stereotyped	['***ingénieur***','*ingénieure*', '***ingénieur.e***']
An engineer works at an industrial [...] new equipment.	***Un.e ingénieur.e*** travaille sur un site industriel [...] équipements.	inclusive	ambiguous	male-stereotyped	['***ingénieur***','*ingénieure*', '***ingénieur.e***']

Figure 2: Examples from the FairTranslate dataset. Words in colors indicate *ambiguous/neutral*, *male*, *female*, and *inclusive* entities.

- **French**: French translations, which are faithful to the gender of the English original. These sentences serve as a ground-truth reference for the English-French translation task.
- **Gender (male / female / inclusive)**: This column indicates the intended grammatical gender of the occupation noun in the French translation.
- **Ambiguity (ambiguous / unambiguous / long unambiguous)**: This column specifies the clarity of the referent's gender in the English original.
  - *ambiguous*: No explicit cues in English indicate the referent's gender, allowing multiple valid interpretations (masculine, feminine, or inclusive) in French.
  - *unambiguous*: Pronouns or other linguistic cues make the referent's gender immediately clear (e.g., '*he*', '*she*' or '*they*' in close proximity to the target noun).
  - *long unambiguous*: The gender can be inferred, but only from contextual cues appearing after a delay (e.g., multiple sentences later), thus testing a model's ability to perform coreference resolution over longer stretches of text.

  By differentiating these categories, we can analyze whether models default to masculine forms in ambiguous cases and whether they can accurately resolve pronouns when explicit signals are present, either nearby or further afield.
- **Stereotype (male-stereotyped / female-stereotyped / gender-balanced)**: Each sentence refers to an occupation assigned to one of these categories, based on real-world gender distribution data[2]. Professions statistically dominated by men (e.g., '*mechanic*'), dominated by women (e.g., '*nurse*') or gender-balanced (e.g., '*attorney*') challenge models to translate consistently regardless of stereotype.
- **Occupation (list of gendered forms in French)**: For each occupation noun, this column contains a list of its three gendered forms in French (male, female, and inclusive). For example, if the English original refers to a '*nurse*', the corresponding value is ['*infirmier*', '*infirmière*', '*infirmier.ière*']. This information enables precise evaluation of models' ability to accomplish grammatical gender agreement across different linguistic forms while maintaining the the intended meaning.

4.2 Dataset Construction

We adopted a two-step approach to generate and annotate the FairTranslate Dataset. First, we designed target sentences in French. This has enabled us to capture all the relevant gender-marking structures inherent to French. Second, we translated these French sentences into English while maintaining consistency across the gender variants.

Step 1: French Sentence Generation and Annotation

- Selecting Occupations (see Figure 3): We begin with three curated lists of occupations:

[2]The gender distribution data for occupation was sourced from Statbel, the official Belgian statistical office.

Feminine

["dietician", "cleaner", "schoolteacher", "childminder", "nursing assistant", "nurse", "pharmacy assistant", "hairstylist", "beautician", "cashier", "teller", "accounting employee", "social worker", "pharmacist", "salesperson", "flight attendant", "childcare worker", "caregiver", "daycare worker", "housekeeper", "secretary", "librarian"]

Masculine

["construction worker", "logger", "firefighter", "electrician", "welder", "plumber", "mechanic", "carpenter", "joiner", "electromechanic", "street sweeper", "garbage collector", "butcher", "engineer", "bus driver", "supervisor", "computer scientist", "programmer", "police officer", "surgeon", "construction machine operator", "upholsterer"]

Neutral

["chemical technician", "management controller", "press operator", "buyer", "quality technician", "physiotherapist", "lawyer", "server", "teacher", "product manager", "translator", "project manager", "career counselor", "doctor", "optician", "special education teacher", "journalist", "accountant"]

Figure 3: Lists of occupations used for sentence generation, grouped by stereotypical gender.

  - Male-stereotyped (22 occupations predominantly held by men)
  - Female-stereotyped (22 occupations predominantly held by women)
  - Gender-balanced (18 occupations held by men and women at similar rates)

  This selection served a dual purpose: it enabled the construction **stereotype** and **occupation** variables, and it produced a diverse set of sentences for testing machine translation across a range of occupational contexts.
- Sentence Creation with GPT Assistance: Using GPT-4o [21] and GPT-o1 [22] under close human supervision, we generated French sentences that captured each occupation and related activities. A human operator prompted and reviewed the model outputs, making manual edits as needed to ensure variety, relevance, and proper structure. This hands-on oversight was critical to guarantee high-quality and contextual validity of the sentences.
- Ambiguity Annotations for English Sentences: For each occupation, we generated three types of sentences:
  - Ambiguous (5 sentences): No explicit pronoun or other cue indicates the person's gender for English translations. French sentences were by nature unambiguous since the form of the occupation is gendered.
  - Unambiguous (5 sentences): A pronoun or nearby context reveals the gender for the English translation.
  - Long Unambiguous (3 sentences): Gender is determinable only after multiple lines or sentences of text, testing longer-range coreference resolution for the English translation.

  These categories allowed us to label each example with the "ambiguity" variable, ensuring clarity regarding when and how gender clues appear in the text.
- Gender Variants in French: Each French sentence was replicated in masculine, feminine, and inclusive forms. We did this by making changes only to the occupational forms and the related satellite elements, while keeping other elements constant across variants. The **Occupation** column was used to systematically retrieve and apply the correct gendered forms for each occupation. This method ensured that the sentences remained semantically identical across the three gender variants.
- Final Checks and Annotation: Human annotators verified that each sentence was correctly annotated across the variables. This process has helped ensure the integrity and consistency of the data prior to translation.

Step 2: English Translation

Once the French sentences and their annotations (gender, ambiguity, stereotype, occupation) were finalized, we used GPT-4o to translate them into English, adhering to the following guidelines:

- Maintaining Consistency Across Gender Variants: For each set of three French sentences (masculine, feminine, inclusive) belonging to the same base example, we produced aligned English translations in which all aspects remained identical except for the explicitly gendered elements (e.g., '*he*', '*she*', or singular '*they*' pronouns; possessive forms '*his*', '*her*', '*their*')[3].
- Human Verification: We tasked a human reviewer with the quality assessment of the automated translations. In particular, the task was to confirm the accuracy of the translation, and ensure that no changes beyond the gender markers had been introduced. This step has helped ensure consistent pairing between the French sentences and their English counterparts.

4.3 The Challenges of Inclusive Language

Inclusive language refers to ways of expressing oneself that aim to ensure women are not excluded in occupational contexts, individuals are addressed appropriately either with respect to or regardless of their gender identity (male, female, or non-binary), and gender diversity is highlighted when referring to mixed groups. These aims adapt to the linguistic characteristics of each language.

In English, proponents of inclusive language have proposed a variety of forms over the years, including neopronouns such as *'xe'* or *'ze'*. However, the singular *'they'*, along with the possessive *'their'* and gender-neutral occupational terms (e.g., *'Chair'* instead of *'Chairman'* or *'Chairwoman'*), has emerged as the most widely accepted standard due to its historical roots and frequent contemporary use. For this reasons, we have opted to use this standard in the construction of our dataset.

In French, inclusive language is far less standardized, where the diverse existing practices have been described as a "graphic tumult" [1]. For occupational nouns, French speakers tend to use lexical

[3]This only applies to possessive forms in English, as in French, possessive forms do not agree with the possessor.

doublets (e.g. '*enseignant*' (male) and '*enseignante*' (female)). Inclusive forms, on the other hand, are represented with a wide variety of typographical signs (*'enseignant.e', 'enseignant/e', 'enseignant-e', 'enseignantE'*). Similarly, inclusive neopronouns (*'iel'/'al'/'ul'*) and determiners (*'lea'/'la·le'/'lo'*) are still subject to great variability. Efforts to systematize these practices include works like Alpheratz [3, 4], Touraille and Allassonnière-Tang [30], Hadad's guidelines[4], and the recommendations of the HCEHF[5]. Other innovative approaches include inclusive typography, as explored by Circlude [7] and the ByeByeBinary collective's type library[6].

In the construction of the FairTranslate dataset, we opted for the following inclusive French writing model. We used a selected form for each occupational term and its related elements. For the singular *'they'*, we opted for French *'iel'*, paring the indefinite English article *'a(n)'* with French *'un.e'*, and the definite *'the'* article with French *'lea'*. Due to the lack of standards, we had to make choices in the construction of FairTranslate. In order to accommodate for other French inclusive writing practices, we developed a Python dictionary made available on GitHub[7]. The dictionary provides an easy way to map our chosen inclusive forms to a wider range of recognized alternatives. The dictionary enables machine translations to be evaluated as correct even if they use a different inclusive form than those explicitly listed in the dataset (e.g., *'ul'* instead of *'iel'*). By offering this adaptable tool, we aim to support inclusive translation research while remaining receptive to the natural evolution of inclusive language.

5 Experimental Setup

As introduced in the previous section, we use the FairTranslate dataset specifically designed to evaluate gender bias in translation. The goal of this section is not to fine-tune or train the models, but to use them in a zero-shot evaluation setting. We pass all English sentences from FairTranslate to each of the selected LLMs and collect their French translations. This allows us to systematically assess how each model handles gendered and inclusive language without any additional training.

For each experiment, we evaluate a range of open-source LLMs with varying sizes and architectures to ensure broad coverage and generalizability of our findings. The selected models include: Gemma2 2B [29]; Mistral 7B [18]; Llama3.1 8B, and Llama3.3 70B (equivalent to Llama3.1 405B) [11].

We introduce several prompting strategies:

- **Task prompting**: We define task prompting as the baseline instruction for the translation task:

  "*Translate the following sentences from English to French:* '{english_sentence}'. *Respond with the translation only, nothing else.*"

The task prompting serves as a simple directive for the model to perform the translation without any additional guidance or constraints. It will be used as a **Baseline** in our experiments. Building on this, we introduced three prompting strategies to encourage gender-inclusive translations.

We hypothesized that the observed gender biases in the generated translations can be attributed to a lack of moral and linguistic awareness, as previously suggested in studies such as Hansen and Żółtak [17] and Zhao et al. [32]. To account for these possibilities, we introduced additional prompting strategies designed to elicit moral and linguistic knowledge:

- **Moral prompting**: To promote gender-inclusive and inclusive translations, we provide the following instruction –proposed by Chen et al. [5]– before the task prompting:
  "*You are a translation model without gender bias and LGBTQA+ friendly.*" + task prompting.
- **Linguistic prompting**: To explicitly encourage the use of gender-inclusive linguistic forms in French, we append the following instruction:
  "*Forms like 'iel' as a neutral pronoun, 'un·e,' 'lea,' or 'ce·tte' as neutral determiners, or a mid-dot (e.g., 'étudiant·e') for gender-neutral terms to be applied only if explicitly requested.*" [8] + task prompting.
- **Moral and Linguistic prompting**: This strategy combines the moral and linguistic promptings to provide both ethical and grammatical guidance for gender-inclusive translations:
  Moral prompting + linguistic prompting + task prompting.

By evaluating translations under these three promptings — *moral prompting*, *linguistic prompting*, and *moral and linguistic prompting* — in addition to the baseline task prompting, we aim to quantify their respective impacts on reducing gender bias.

6 General Translation Performance

This section evaluates the general translation performance of four leading Large Language Models (LLMs) on the FairTranslate dataset. Our primary objective is to assess the translation quality across gender categories (male, female, and inclusive) using two complementary metrics: BLEU and COMET. These metrics provide a dual perspective on translation accuracy, balancing surface-level quality and deeper semantic correctness. By analyzing the results, we aim to identify patterns in how LLMs handle translations involving gendered and inclusive language, as well as disparities between gender categories.

6.1 Translation Evaluation Metrics

We adopt two widely recognized metrics to evaluate translation quality. **BLEU** [23] measures n-gram overlap between machine-generated and reference translations. While it provides a quick and interpretable measure of translation quality, it is limited in capturing semantic nuances. In contrast, **COMET** [26], specifically the wmt22-comet-da model [25], is a neural-based metric fine-tuned on human-annotated datasets to assess semantic alignment and translation quality. COMET captures subtle linguistic variations, including gender-related adaptations, and demonstrates stronger correlations with human judgment compared to BLEU. The use

[4] https://www.motscles.net/ecriture-inclusive
[5] https://www.haut-conseil-egalite.gouv.fr/IMG/pdf/guide_egacom_sans_stereotypes-2022-versionpublique-min-2.pdf
[6] https://typotheque.genderfluid.space/fr
[7] https://github.com/fanny-jourdan/FairTranslate

[8] In addition to the context linguisitc, we were obliged to add this sentence: "*Otherwise, use the classic feminine or masculine form.*" to counter the LLMs' bias of putting all sentences in neutral because the prompt was talking about it.

Model	Gemma2-2B		Mistral-7B		Llama3.1-8B		Llama3.3-70B		Mean	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Female	42.34	87.86	38.94	87.16	43.73	88.88	51.44	90.12	44.16	88.56
Male	**45.37**	**89.18**	**41.36**	**88.25**	**46.18**	**89.85**	**55.86**	**91.35**	**47.19**	**89.66**
Inclusive	36.65	85.42	33.77	84.42	37.53	85.97	45.27	87.16	38.35	85.80

Table 1: Comparison of BLEU and COMET scores (in percentages) for different genders and models, with mean scores across models for Baseline Prompting. The bold font indicates the gender for which the model performs best. We observe that models show best performance on the male gender, and worst performance on the inclusive gender. This finding does not vary across models or translation quality metrics.

of both metrics enables a balanced evaluation and allows for the analysis of translation performance from multiple dimensions.

6.2 Global Results

Table 1 reports BLEU and COMET scores for male, female, and inclusive gender categories across the four LLMs. Across all models, scores are highest for the male gender, followed by the female gender, and lowest for the inclusive gender. ANOVA tests (Table 6) indicate that these differences are statistically significant, highlighting a systemic disparity in translation quality. As each dataset example exists in all three gender variants, the observed differences can be attributed to the model performance rather than to the dataset bias. These findings emphasize persistent challenges in achieving gender-inclusive translation, with LLMs consistently underperforming on translation of gender inclusive forms.

6.3 Effects of Prompting

Table 2 evaluates the translation performance of four prompting strategies (baseline, moral, linguistic, and moral+linguistic) as indicated by BLEU scores across gender categories. Results using COMET scores are provided in the Appendix C.

For the male and female genders, translation performance generally declines as more information is incorporated into the prompts. For the inclusive gender, we observed performance decreases for Mistral and performance improvements for Gemma2, Llama3.1, and Llama3.3 models. Overall, translation performance tends to decrease for the traditional binary genders (male and female) and increase for the inclusive gender as the prompting strategies become more informative. These findings largely reflect the fact that our prompting strategies were designed the improve the translation performance of inclusive gender forms. The same prompts which are successful at that task appear to create noise for the translation of traditional gender forms. Notably, while the performance on both female and male categories declined with prompting, the decline was steeper for male gender. This had the effect offurther narrowing the performance gap which was initially observed between the male and female gender translations. Consequently, the overall disparity among all gender categories is reduced. In spite of this reduction, the male category still had better translation results, at statistically significant rates. This indicates that complete equity in machine translation has not yet been achieved.

6.4 Analysis of Metadata Variables

To analyze biases in more detail, we examine the effects of occupational stereotypes and sentence ambiguity on translation performance for Gemma2-2B.

Table 3 shows the performance of Gemma2 across gender categories (male, female, inclusive) and two key variables: *Stereotype* and *Ambiguity*. The left facet reports BLEU and COMET scores for sentences categorized by occupational stereotypes (female-stereotyped, male-stereotyped, gender-balanced). The right facet provides a breakdown of the ambiguity categories (ambiguous, long unambiguous, unambiguous). Results for other models, provided in Appendix D, demonstrate consistent behavior across these variables. Across all stereotype and ambiguity categories, the instances with male gender were consistently better translated than those with other genders, irrespective of the variable examined.

For stereotype categories, even for female-stereotyped occupations, male translations outperformed female translations. While the gap between male and female scores was the smallest in this category, less than 2% in BLEU and less than 1% in COMET, it is notable that female gender translations did not achieve higher scores, as one might expect given the stereotypical association. This could be due to the historical tendency in French to default to the masculine form for occupations, even when those occupations are predominantly associated with women. Unsurprisingly, male-stereotyped and gender-balanced occupations showed a larger gap. In all cases, scores for the inclusive gender were lower than those for male and female genders, with the largest differences observed in gender-balanced occupations (11% in BLEU and 4% in COMET). The generally superior performance in gender-balanced occupations may reflect greater diversity of gender representation within these examples in training data.

For ambiguity categories, BLEU scores were highest for long unambiguous sentences (49.16%), followed by unambiguous (40.51%) and then ambiguous (37.71%) sentences. However, these differences likely reflect a bias in BLEU rather than a true model effect. As shown in Table 5, sentences in the long unambiguous category were significantly longer (48.6 words on average) than those in unambiguous (22.2 words) or ambiguous (15.8 words) categories. BLEU score is highly sensitive to sentence length: errors in longer sentences exert less impact on the score than errors in shorter ones. In contrast, COMET results indicate higher scores for ambiguous sentences. This difference may stem from COMET's robustness to lexical variations, particularly when predicting gendered forms of

Prompting	Gemma2-2B			Mistral-7B			Llama3.1-8B			Llama3.3-70B		
	Female	Male	Inclusive	Fem.	Male	Incl.	Fem.	Male	Incl.	Fem.	Male	Incl.
Baseline	**42.34**	**45.37**	36.65	**38.94**	**41.36**	**33.77**	**43.73**	**46.18**	37.53	**51.44**	**55.86**	45.27
Moral	42.04	45.03	37.33	37.43	40.62	33.12	42.03	45.49	37.76	45.02	50.40	43.87
Linguistic	42.11	44.39	**37.70**	36.78	40.81	33.48	42.96	44.68	36.98	51.40	55.32	45.28
Moral & Ling.	42.11	44.70	37.08	36.36	39.29	32.22	42.82	45.24	**38.26**	50.57	55.27	**45.94**

Table 2: BLEU scores for different promptings and genders. The bold font indicates the prompting strategy which performs best for each model-gender combination.

Stereotype	Gender	Gemma2-2B	
		BLEU	COMET
Female-stereotyped	Female	40.78	87.84
	Male	**42.35**	**88.33**
	Inclusive	34.88	84.60
	Mean	*39.34*	*86.92*
Male-stereotyped	Female	41.33	87.13
	Male	**44.69**	**88.76**
	Inclusive	36.45	85.42
	Mean	*40.82*	*87.10*
Gender-balanced	Female	45.29	88.72
	Male	**49.63**	**90.63**
	Inclusive	38.91	86.36
	Mean	***44.61***	***88.57***

Ambiguity	Gender	Gemma2-2B	
		BLEU	COMET
Ambiguous	Female	36.65	88.25
	Male	**41.32**	**89.36**
	Inclusive	35.17	85.78
	Mean	*37.71*	***87.80***
Long Unambiguous	Female	50.35	87.66
	Male	**51.66**	**88.77**
	Inclusive	45.48	86.67
	Mean	***49.16***	*87.69*
Unambiguous	Female	43.17	87.60
	Male	**45.61**	**89.25**
	Inclusive	32.76	84.30
	Mean	*40.51*	*87.05*

Table 3: BLEU and COMET scores for the Gemma2-2B model and each combination of (Top) stereotype and (Bottom) ambiguity with gender. Results are obtained with Baseline Prompting. Bold font highlights the best-performing gender per group. Bold italicized font shows the best-performing stereotype or ambiguity category. Additional results appear in Appendix D.

professions (e.g., '*chirurgien*' vs. '*chirurgienne*'), which BLEU treats as distinct tokens.

The BLEU score has consistently exhibited larger gaps both between genders, and across variables. We ascribe this to to its sensitivity to surface-level changes, such as the exact form of an occupation. COMET, being less affected by such lexical variations, provides a more nuanced measure of translation quality. However, COMET's robustness to gendered differences can sometimes obscure disparities in model performance, making it less sensitive to errors related to gender agreement.

7 Specific Analysis of Occupational Terms

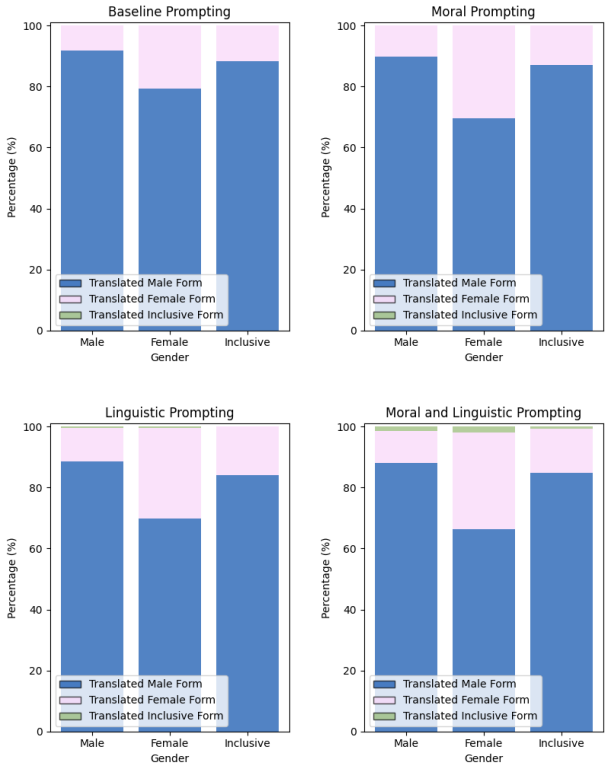

Figure 4: Comparison of the Labelled Gender Based on the Form of the Translated Occupation by Gemma2-2B for Different Promptings.

Models	Baseline	Moral Prompting	Linguistic Prompting	Moral and Linguistic Prompting
Gemma2-2B	5	9	29	49
Mistral-7B	0	1	1	2
Llama3.1-8B	1	4	10	42
Llama3.3-70B	0	19	12	86

Table 4: Number of gender-inclusive french translated sentences detected using gender-inclusive indicators in different models in different promptings out of 806 real gender-inclusive sentences in FairTranslate.

Accurately translating gendered forms of occupations in French addresses two key challenges: ensuring feminine forms are used for traditionally male-dominated occupations and incorporating modern inclusive forms to represent individuals beyond the binary gender spectrum. Both are critical for fairness and inclusivity in machine translation.

This experiment evaluates the Gemma2-2B model (other models are analyzed in appendix E) in translating occupations from English to French, focusing on how gender is represented in the translated forms (as explained in Section 2, French offers three gender grammatical forms). The goal is to assess whether the grammatical gender of the translated occupation aligns with the actual gender associated with the original sentence.

The dataset is filtered to retain only sentences where the correct occupation is translated by the model. For example, if the English sentence contains *'nurse'* and the model translates it as *'chirurgien'/'chirurgienne'/'chirurgien.ne'* (surgeon) instead of *'infirmier'/'infirmière'/'infirmier.ère'* (nurse), the sentence is excluded. The grammatical gender of the translated occupation is automatically identified using a predefined mapping, and it is compared to the original sentence's gender label. For instance, if the original sentence referred to a female nurse, the expected gender is feminine, and we check whether the translation uses the form *'infirmière'*. This comparison allows us to measure how often the translation respects the intended gender.

Figure 4 highlights a significant bias: the model overwhelmingly defaults to the masculine form, regardless of whether the original subject is male, female, or inclusive. Inclusive forms (e.g., infirmier.ière) are almost entirely absent, reflecting the model's inability to adopt recent linguistic innovations for inclusive representation. Attempts to improve inclusivity, such as "Moral Prompting" or "Linguistic Prompting," fail to resolve the issue and can even exacerbate it in some cases (see Appendix E).

8 Specific Analysis of Inclusive Gender

In Section 6, we showed that models perform significantly worse when translating inclusive gender forms compared to masculine and feminine forms across all configurations. Additionally, in Section 7, we observed that occupational terms are almost never translated into inclusive forms in French. Here, we extend the analysis to all types of inclusive gender indicators.

8.1 Inclusive Gender Indicators

We first investigate whether the poor translation of occupational terms into inclusive forms reflects a broader difficulty in translating inclusive gender indicators overall. Table 4 presents the frequency of inclusive gender indicators appearing in French translations for various models and prompting strategies.

Inclusive gender indicators include terms such as ["Iel", "iel", "Lea", "lea", "Un.e", "un.e", "Ce.tte", "ce.tte"], as well as inclusive occupational forms identified by endings like ["ien.ne", "ier.ère", "eur.euse", "eur.e", "eux.euse", "tre.esse", "te.esse", "eur.rice"]. These indicators are assessed across 806 sentences annotated as inclusive in the gender column of the dataset. Each of these sentences in the FairTranslate dataset includes at least one inclusive indicator or occupational form in its correct French translation.

However, the models rarely produce the expected inclusive indicators or forms. Across all configurations, the number of sentences containing correctly translated inclusive forms ranges between 0 and 86 out of 806, which represents less than 11% of sentences in the best case.

Although this performance is insufficient, we observe a trend: the use of moral or linguistic prompting improves the translation of inclusive forms compared to the baseline. Furthermore, combining both types of prompting provides the best results for all models tested.

8.2 Inclusive They Translation Analysis

To understand the errors models make when translating inclusive gender, we conduct a focused analysis of the singular '*they*'. If models fail to translate inclusive indicators, what do they produce instead? For this analysis, we focus on the singular '*they*' and examine 246 examples where it appears explicitly in the English source. Cases involving other inclusive markers (e.g., '*their*') or ambiguities visible only in the French translation are excluded.

The singular '*they*' can be translated into French in various ways by a model, but not all of which are correct or appropriate:

- **Plural forms:** such as '*ils*' (masculine plural), '*elles*' (feminine plural) or '*iels*' (inclusive plural). This suggests the model has not recognized the singular usage of '*they*'.
- **Singular gendered forms:** such as '*il*' (masculine singular) or '*elle*' (feminine singular). This suggests the model understands the singular usage of '*they*' but does not account for inclusive gender practices in French.
- **Inclusive forms:** such as '*iel*', which reflects correct alignment with recent inclusive practices in French.
- **Pronoun omission:** a valid strategy in French for achieving inclusivity without explicitly using a pronoun.

Figure 5 shows that Gemma2-2B predominantly translates singular '*they*' as a plurial. Prompting strategies improve the use of inclusive forms (e.g., '*iel*' or omission) but remain insufficient, with plural translations still occurring in nearly 50% of cases. Appendix

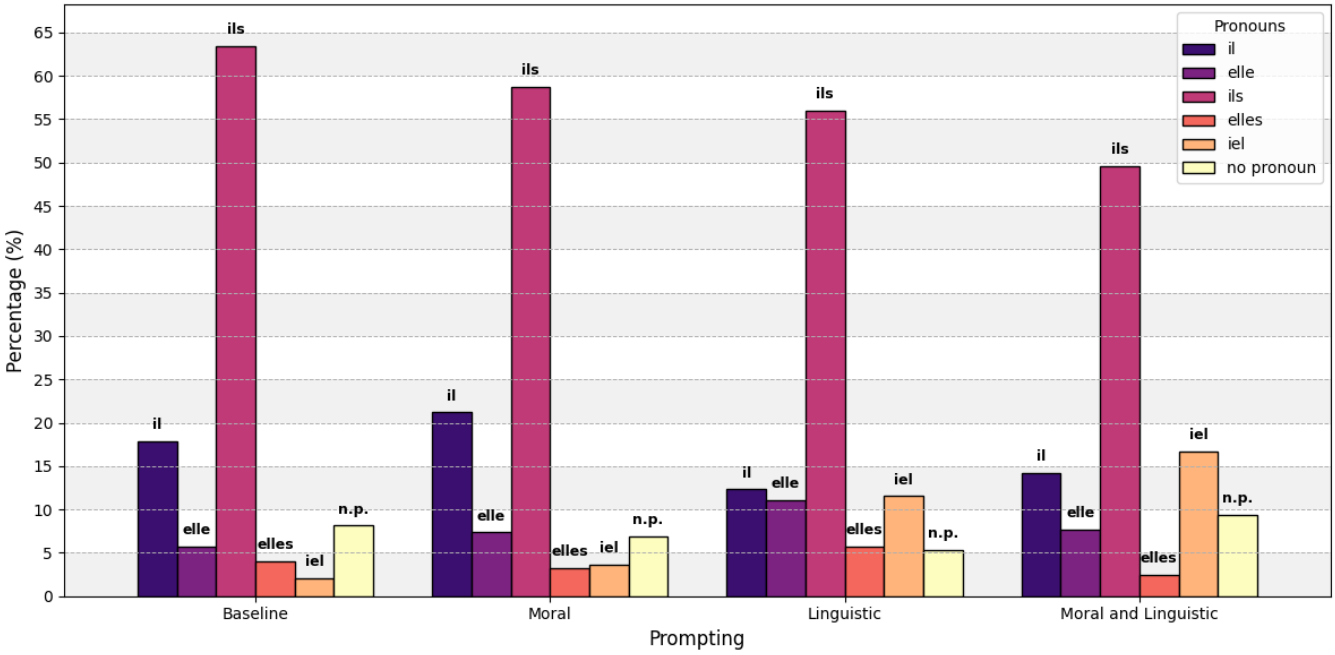

Figure 5: Distribution of French Pronouns in the Translation of the Inclusive *They* with Gemma2. The figure is based on 246 English sentences containing the pronoun '*they*' labeled as 'inclusive' and shows the translation results under the four promptings. The predominance of '*ils*' in the figure highlights a bias in the model, which struggles to fully capture the inclusive aspect of '*they*'.

F provides similar results for other models, often showing even worse performance.

These results reveal that the issue extends beyond models failing to adopt the relatively new inclusive practices in French. The problem is deeper: models often fail to correctly interpret inclusive usages in English, such as the singular '*they*', which they misinterpret as plural. This fundamental misunderstanding highlights that the challenge is not merely one of cultural or linguistic adaptation to French but also reflects a lack of comprehension of long-established inclusive constructs in English. Addressing this requires substantial improvements in how models process and represent nuanced linguistic phenomena.

9 Conclusion

In this work, we introduced FairTranslate, a novel English-French translation dataset annotated by experts with extensive metadata, enabling detailed analysis of non-binary gender biases in translation. This dataset is a valuable resource for evaluating machine translation systems, particularly in handling inclusive gender forms. Additionally, we evaluated four LLMs using varied prompting strategies which has offered novel insights into how these models process gender inclusivity in translation tasks.

Our findings reveal consistent shortcomings in the translation of inclusive gender forms. Across all configurations, models performed significantly worse on inclusive gender translations compared to masculine and feminine forms. Moreover, French inclusive gender indicators and forms were almost never used, even when prompting was applied. This underscored the models' inability to integrate relatively recent linguistic practices in French.

Crucially, our analysis demonstrates that the inclusive language challenges extend beyond challenges which are specific to French. The poor translation of inclusive gender forms is rooted in a more fundamental problem: models fail to adequately understand inclusive constructs in English, such as the singular '*they*'. The misinterpretation of these constructs as plural contributes to their inadequate French translations. This finding highlights the need for LLMs to better capture the nuances of inclusive language, both in English as well as in other languages such as French.

By publicly sharing our dataset and analysis, we aim to encourage the development of equitable translation systems that are not only linguistically competent but also inclusive, addressing biases that disproportionately affect underrepresented linguistic forms.

Acknowledgments

The authors thank all the people and industrial partners involved in the FOR and DEEL projects. This work has benefited from the support of the FOR[9] and DEEL[10] projects, with fundings from the

[8]Note that in some cases, the use of the plural may be justified, as the singular '*they*' of some English examples can lead to a double interpretation. However, this is not a majority of the examples in the dataset, so there is a real bias towards over-use of this translation.

[9]https://www.irt-saintexupery.com/fr/for-program/

[10]https://www.deel.ai/

Agence Nationale de la Recherche, and which is part of the ANITI AI cluster.

A special thank you goes to Daniel Anadria for his help and insightful suggestions throughout the writing process. His careful proofreading and constructive feedback greatly enhanced the clarity and overall quality of this manuscript.

References

[1] Julie Abbou. 2013. Pratiques graphiques du genre. *Langues et cité* 24 (2013), 4–5.
[2] Anne Abeillé and Danièle Godard. 2021. *La grande grammaire du français*. Éditions Actes Sud.
[3] Alpheratz. 2018. *Grammaire du français inclusif: littérature, philologie, linguistique*. Vent solars. https://books.google.fr/books?id=z1fmvQEACAAJ
[4] Alpheratz. 2019. Français inclusif : du discours à la langue ? *Le Discours et la Langue Revue de linguistique française et d'analyse du discours* 111 (2019), 53–74. https://hal.science/hal-02323626
[5] Yijie Chen, Yijin Liu, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2024. Beyond Binary Gender: Evaluating Gender-Inclusive Machine Translation with Ambiguous Attitude Words. *arXiv preprint arXiv:2407.16266* (2024).
[6] Yannick Chevalier, Hughes Constantin de Chanay, and Laure Gardelle. 2017. Bases linguistiques de l'émancipation: système anglais, système français. *Mots. Les langages du politique* (2017), 9–36.
[7] Camille Circlude. 2023. La typographie post-binaire. Au-delà de l'écriture inclusive. *Paris, Editions* 42 (2023), 224.
[8] Greville G Corbett. 1991. *Gender*. Cambridge University Press.
[9] Anna Currey, Maria Nădejde, Raghavendra Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*.
[10] Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. In *Queer in AI Workshop at ACL 2023*.
[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
[12] Daniel Elmiger. 2017. Binarité du genre grammatical–binarité des écritures? *Mots. Les langages du politique* 113 (2017), 37–52.
[13] Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL 2023)*.
[14] Vagrant Gautam, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow. 2024. WinoPron: Revisiting English Winogender Schemas for Consistency, Coverage, and Grammatical Case. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, Maciej Ogrodniczuk, Anna Nedoluzhko, Massimo Poesio, Sameer Pradhan, and Vincent Ng (Eds.). Association for Computational Linguistics, Miami, 52–66. https://doi.org/10.18653/v1/2024.crac-1.6
[15] Pascal Gygax, Ute Gabriel, Arik Lévy, Eva Pool, Marjorie Grivel, and Elena Pedrazzini. 2012. The masculine form and its competing interpretations in French: When linking grammatically masculine role names to female referents is difficult. *Journal of Cognitive Psychology* 24, 04 (2012), 395–408.
[16] Pascal Gygax, Sandrine Zufferey, and Ute Gabriel. 2021. Le cerveau pense-t-il au masculin. *Cerveau, langage et représentations sexistes, Paris, Le Robert* (2021).
[17] Karolina Hansen and Katarzyna Żółtak. 2022. Social perception of non-binary individuals. *Archives of Sexual Behavior* 51, 4 (2022), 2027–2035.
[18] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
[19] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*. 12–24.
[20] Manuel Lardelli, Giuseppe Attanasio, and Anne Lauscher. 2024. Building Bridges: A Dataset for Evaluating Gender-Fair Machine Translation into German. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7542–7550. https://doi.org/10.18653/v1/2024.findings-acl.448
[21] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.
[22] OpenAI. 2024. Introducing OpenAI o1-preview: A new series of reasoning models for solving hard problems. https://openai.com/index/introducing-openai-o1-preview/.
[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135
[24] Manuel Pérez, Katy Barasc, and Hélène Giraudo. 2019. Des (dés) accords grammaticaux dans la dénomination écrite de la personne en France: un tumulte graphique entre passions tristes et passions joyeuses. *GLAD!. Revue sur le langage, le genre, les sexualités* 07 (2019).
[25] Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022.

COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*. 578–585.
[26] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213
[27] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301* (2018).
[28] Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL 2019)*.
[29] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).
[30] Priscille Touraille and Arc Allassonnière-Tang. 2023. Chapitre 8. Idéer une catégorie épicène et la matérialiser cohéremment dans la langue. In *Qu'est-ce qu'une femme?* Éditions Matériologiques, 167–233.
[31] Andreas Waldis, Joel Birrer, Anne Lauscher, and Iryna Gurevych. 2024. The Lou Dataset - Exploring the Impact of Gender-Fair Language in German Text Classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10604–10624. https://doi.org/10.18653/v1/2024.emnlp-main.592
[32] Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. 2021. Ethical-advice taker: Do language models understand natural language interventions? *arXiv preprint arXiv:2106.01465* (2021).
[33] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2018)*.

A Average Number of Words per Example in FairTranslate

Table 5 shows the average number of words (in both English and French) per example in the FairTranslate dataset, categorized by gender ambiguity type. Sentences labeled as "ambiguous" are the shortest, with an average of 15.7 words in French and 13.6 in English, as they contain no gender information in English. These are followed by sentences labeled as "unambiguous," which average 22.2 words in French and 19.1 in English. The longest sentences are those labeled as "long unambiguous," designed to include a gender indicator in English placed as far as possible from the profession (which is gendered in the French translation) to study long-distance coreference resolution. These sentences average 48.6 words in French and 40.1 in English. It is worth noting that French sentences are generally longer than their English counterparts due to structural differences between the two languages.

Ambiguity	French	English
Ambiguous	15.775	13.564
Unambiguous	22.153	19.102
Long Unambiguous	48.599	40.104
#Avg Word Count	**28.842**	**24.257**

Table 5: Average Word Count for French and English Sentences by Ambiguity Category in FairTranslate Dataset.

B ANOVA tests

Table 6 presents the p-values from all ANOVA tests conducted to verify whether the differences in scores (for BLEU and COMET) across genders are statistically significant. All p-values are well below the 0.05 threshold, with every value being smaller than 10^{-10}.

C Effects of Prompting for COMET scores

Table 7 evaluates the effect of four prompting strategies—baseline, moral, linguistic, and moral+linguistic—on translation performance, focusing on COMET scores across gender categories.

- Male Gender: Performance generally decreases as prompting strategies incorporate more information (only one exception: Mistral-7B, which shows a slight improvement under the moral prompting).
- Female Gender: Performance also generally decreases (two exceptions: Gemma2-2B with Moral&Linguistic and Llama3.3-70B with Linguistic).
- Inclusive Gender: Performance decreases for Gemma and Mistral, but increases for Llama3.1 and Llama3.3 with linguistic and combined prompts (except under the Moral prompting for Llama3.3).

While overall performance tends to decline across all genders (except for inclusive gender on 2 models) with increased prompting, the decline is most pronounced for the male gender, reducing disparities between genders. However, these reductions remain insufficient for equitable translation performance.

D Analysis of Metadata Variables for All Models

Tables 8 and 9 report the performance of all models across gender categories with respect to the stereotype and ambiguity variables, respectively. The results from these tables align closely with those observed for Gemma2-2B in Table 3. This consistency confirms that the analysis presented in Section 6.4 generalizes well across all models in our experimental setup.

Prompting	Gemma2-2B		Mistral-7B		Llama3.1-8B		Llama3.3-70B	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Baseline	8.9×10^{-20}	2.8×10^{-46}	7.6×10^{-16}	1.6×10^{-40}	1.2×10^{-19}	1.5×10^{-67}	3.2×10^{-25}	2.6×10^{-91}
Moral	1.9×10^{-15}	8.6×10^{-45}	3.9×10^{-15}	2.6×10^{-46}	7.1×10^{-15}	6.1×10^{-59}	4.1×10^{-11}	5.6×10^{-18}
Linguistic	1.6×10^{-11}	1.4×10^{-34}	1.8×10^{-14}	1.1×10^{-35}	6.7×10^{-16}	1.5×10^{-48}	6.1×10^{-23}	3.6×10^{-78}
Moral & Ling.	9.6×10^{-15}	6.5×10^{-38}	5.4×10^{-14}	5.6×10^{-36}	8.4×10^{-13}	3.1×10^{-39}	2.3×10^{-20}	1.8×10^{-74}

Table 6: P-values from ANOVA tests across gender categories (female, male, inclusive) for BLEU and COMET scores, on different models and prompting configurations.

Prompting	**Gemma2-2B**			**Mistral-7B**			**Llama3.1-8B**			**Llama3.3-70B**		
	Female	Male	Inclusive	Fem.	Male	Incl.	Fem.	Male	Incl.	Fem.	Male	Incl.
Baseline	87.86	**89.18**	**85.42**	**87.16**	88.25	**84.42**	**88.88**	**89.85**	85.97	90.12	**91.35**	87.16
Moral	87.80	89.16	**85.42**	86.82	**88.41**	84.30	88.45	89.83	86.15	87.95	89.20	86.52
Linguistic	87.59	88.81	85.39	86.18	87.92	84.03	88.86	89.39	86.16	**90.19**	91.26	87.59
Moral & Ling.	**87.96**	88.78	85.32	86.43	87.93	84.17	88.68	89.20	**86.23**	89.85	91.26	**87.63**

Table 7: COMET scores for different promptings and genders. The bold font means the Prompting performs best for the gender and the model.

Stereotype	Gender	Gemma2-2B		Mistral-7B		Llama3.1-8B		Llama3.3-70B		Mean	
		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Female-stereotyped	Female	40.78	87.84	37.33	86.79	**43.15**	**89.06**	48.79	90.04	42.51	88.43
	Male	**42.35**	**88.33**	**38.91**	**86.89**	42.64	88.88	**50.54**	**90.17**	**43.61**	**88.57**
	Inclusive	34.88	84.60	32.56	82.95	35.45	85.11	42.18	85.84	36.27	84.63
	Mean	*39.34*	*86.92*	*36.27*	*85.54*	*40.41*	*87.68*	*47.17*	*88.68*	*40.80*	*87.20*
Male-stereotyped	Female	41.33	87.13	38.46	86.70	42.70	88.15	52.21	89.92	43.68	87.98
	Male	**44.69**	**88.76**	**41.20**	**88.45**	**45.71**	**89.76**	**57.52**	**91.85**	**47.28**	**89.70**
	Inclusive	36.45	85.42	33.73	85.21	37.91	86.43	46.64	88.05	38.68	86.28
	Mean	*40.82*	*87.10*	*37.80*	*86.79*	*42.11*	*88.11*	*52.12*	*89.94*	*43.21*	*88.49*
Gender-balanced	Female	45.29	88.72	41.34	88.12	45.56	89.48	53.63	90.44	46.45	89.19
	Male	**49.63**	**90.63**	**44.35**	**89.58**	**50.80**	**91.08**	**60.15**	**92.15**	**51.23**	**90.86**
	Inclusive	38.91	86.36	35.20	85.23	39.48	86.45	47.26	87.68	40.21	86.43
	Mean	***44.61***	***88.57***	***40.30***	***87.64***	***45.28***	***89.00***	***53.01***	***90.09***	***45.30***	***88.82***

Table 8: BLUE and COMET scores for each model and each combination of stereotype and gender on Baseline Prompting, with averages per model and per stereotype. The bold font means the gender performs best for the model. Bold italicized font means that the class in stereotype is the best-performing class for the model.

E Specific Analysis of Occupational Terms for All Models

Figures 6, 7, and 8 illustrate the percentage of occupations translated into their masculine (blue), feminine (pink), and inclusive (green) forms for sentences labeled with masculine, feminine, and inclusive occupational forms in the source language. The analysis is presented for three models: Mistral-7B, Llama3.1-8B, and Llama3.3-70B.

For Mistral-7B and Llama3.3-70B, prompting exacerbates the issue. Not only do these models fail to produce any inclusive translations of occupations, but the percentage of occupations translated into the feminine form also decreases.

In contrast, Llama3.1-8B behaves similarly to Gemma2-2B, showing a slight improvement in the percentage of occupations in the feminine form being correctly translated and a marginal introduction of inclusive forms. However, these results remain far from satisfactory, highlighting significant gaps in gender representation in the translations.

Ambiguity	Gender	Gemma2-2B		Mistral-7B		Llama3.1-8B		Llama3.3-70B		Mean	
		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Ambiguous	Female	36.65	88.25	34.09	87.69	38.69	89.08	44.45	90.40	38.47	88.85
	Male	**41.32**	**89.36**	**38.04**	**88.85**	**42.98**	**90.47**	**50.69**	**91.75**	**43.26**	**90.11**
	Inclusive	35.17	85.78	33.65	85.18	36.07	86.77	44.32	87.75	37.30	86.37
	Mean	*37.71*	***87.80***	*35.26*	***87.24***	*39.25*	***88.77***	*46.49*	***89.97***	*39.68*	***88.44***
Long Unambiguous	Female	50.35	87.66	46.31	86.70	51.02	88.46	60.26	89.28	51.99	88.02
	Male	**51.66**	**88.77**	**47.76**	**87.33**	**51.12**	**88.74**	**62.15**	**90.21**	**53.17**	**88.76**
	Inclusive	45.48	86.67	42.79	85.80	46.53	87.10	55.08	88.00	47.47	86.89
	Mean	***49.16***	*87.69*	***45.62***	*86.61*	***49.56***	*88.10*	***59.16***	*89.16*	***50.87***	*87.89*
Unambiguous	Female	43.17	87.60	39.30	86.92	44.33	88.92	53.06	90.35	44.97	88.45
	Male	**45.61**	**89.25**	**40.79**	**88.19**	**46.39**	**89.91**	**57.22**	**91.64**	**47.50**	**89.75**
	Inclusive	32.76	84.30	28.38	82.83	33.51	84.49	40.24	86.06	33.72	84.42
	Mean	*40.51*	*87.05*	*36.16*	*85.98*	*41.41*	*87.77*	*50.17*	*89.35*	*42.06*	*87.54*

Table 9: Mean BLEU and COMET scores (in percentage) for each model and each combination of ambiguity and gender. The bold font means the gender performs best for the model. Bold italicized font means that the class in ambiguity is the best-performing class for the model.

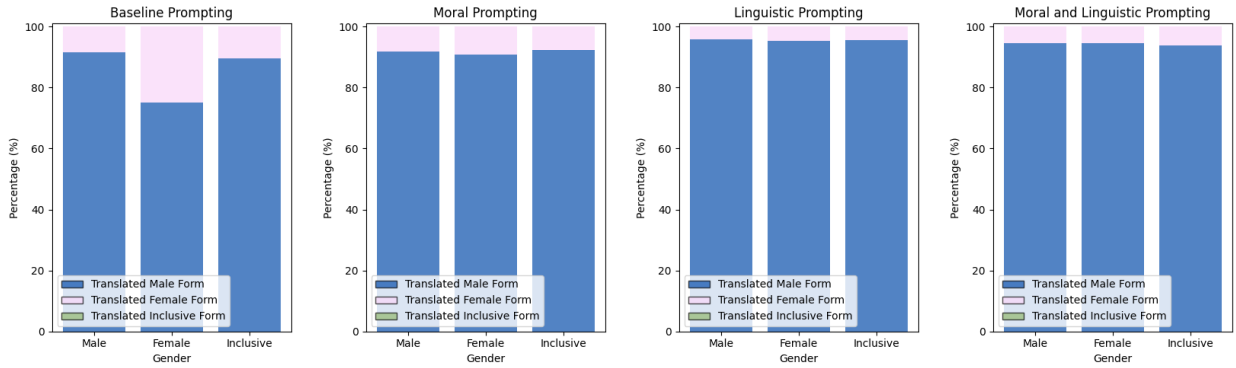

Figure 6: Comparison of the Labelled Gender Based on the Form of the Translated Occupation by Mistral-7B for Different Promptings.

F Inclusive They Translation Analysis for All Models

Figures 9, 10, 11 and 12 present the distribution of French pronouns in the translation of the singular '*they*', filtered by the ambiguity variable. Only the "long unambiguous" and "unambiguous" cases are considered, as "ambiguous" cases are excluded since the singular '*they*' cannot appear in sentences without any gender indicators in English. This analysis evaluates whether context length affects the accurate translation of singular '*they*'.

For Gemma2-2B (Figure 9), the results show that in "long unambiguous" cases, where there is a long context between the occupation (singular) and '*they*', the model translates '*they*' as a plural pronoun over 80% of the time. In contrast, when '*they*' appears close to the singular occupation ("unambiguous"), the plural translation rate drops to 50%, indicating that the model is better able to resolve the link in shorter contexts.

This pattern is much weaker for Mistral-7B (Figure 10) and Llama3.1-8B (Figure 11), which nearly always translate '*they*' as a plural pronoun, regardless of context length or prompting strategy.

For Llama3.3-70B (Figure 12), the baseline and moral prompting strategies yield results similar to Mistral-7B and Llama3.1-8B. However, in "unambiguous" cases (short contexts), applying a linguistic prompting strategy causes the model to translate '*they*' as a singular pronoun, albeit with gendered markers like '*il*' or '*elle*.' This represents a first step toward accurate translation. When both linguistic and moral context prompts are combined, the model more often successfully produces the correct translation of the inclusive singular, '*iel*'. But the more complex case with a long context ("long unambiguous") is still just as bad.

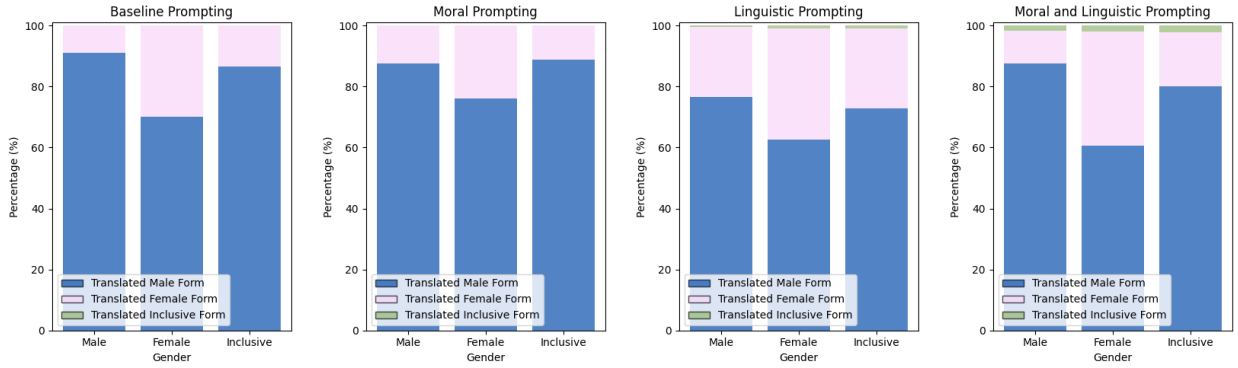

Figure 7: Comparison of the Labelled Gender Based on the Form of the Translated Occupation by Llama3.1-8B for Different Promptings.

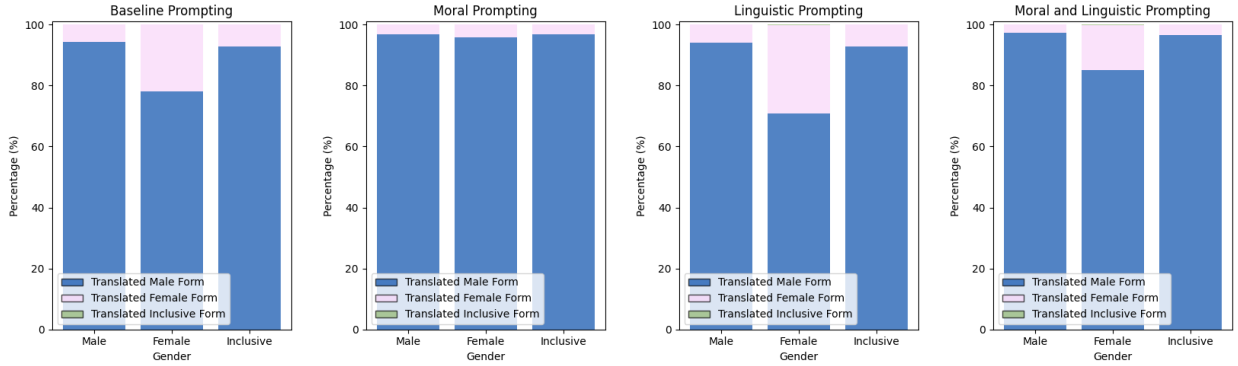

Figure 8: Comparison of the Labelled Gender Based on the Form of the Translated Occupation by Llama3.3-70B for Different Promptings.

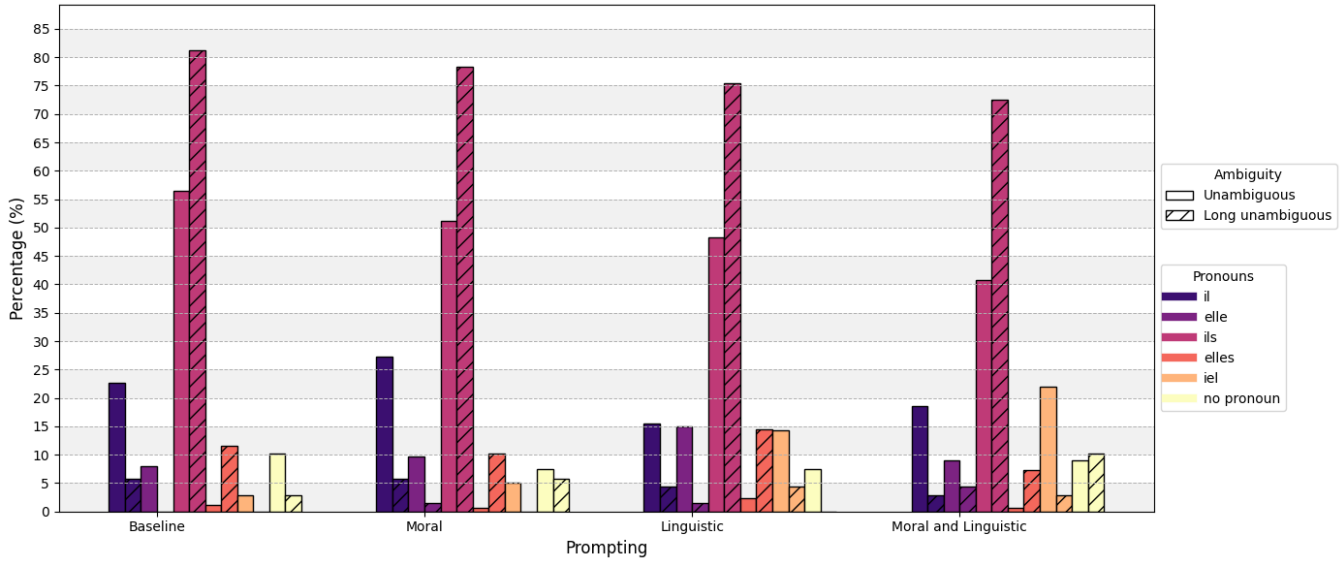

Figure 9: Distribution of French Pronouns in the Translation of the Inclusive *They* with Gemma2-2B by ambiguity variable. The figure is based on 177 unambiguous and 69 long unambiguous English sentences containing the pronoun '*they*' and shows the translation results under the four promptings. The appropriate translations for the inclusive '*they*' are typically either '*iel*' or constructing the sentence without a pronoun (*no pronom*). The predominance of '*ils*' in the figure highlights a bias in the model, which struggles to fully capture the inclusive aspect of '*they*'.

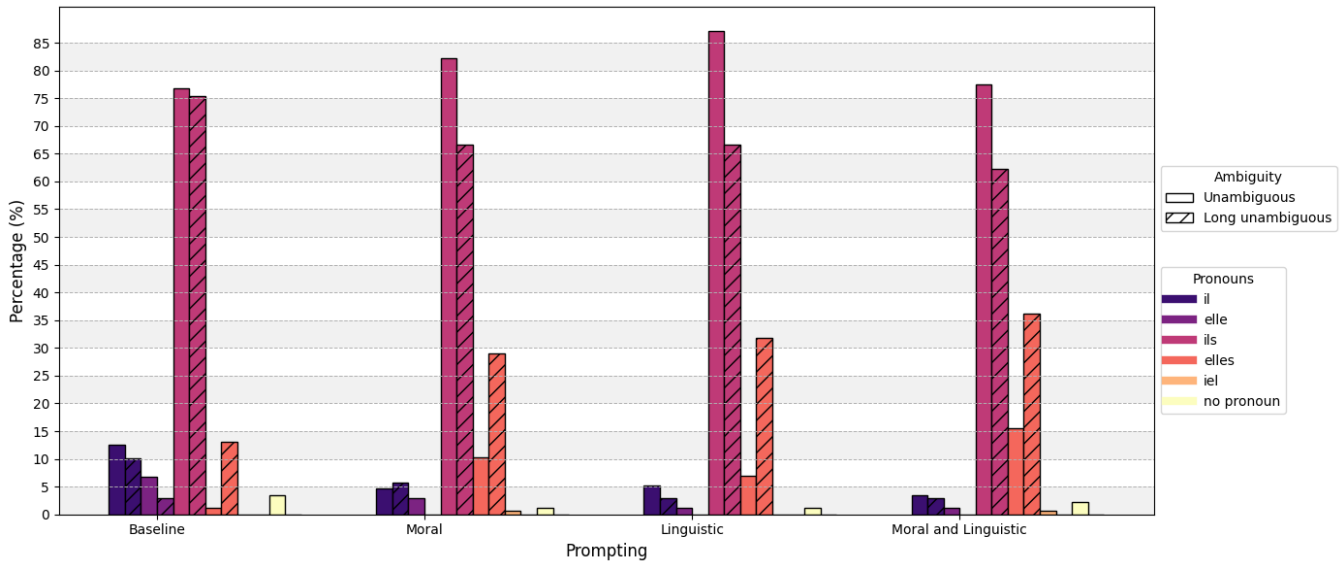

Figure 10: Distribution of French Pronouns in the Translation of the Inclusive *They* with Mistral-7B by ambiguity variable. The figure is based on 177 unambiguous and 69 long unambiguous English sentences containing the pronoun '*they*' and shows the translation results under the four promptings. The appropriate translations for the inclusive '*they*' are typically either '*iel*' or constructing the sentence without a pronoun (*no pronom*). The predominance of '*ils*' in the figure highlights a bias in the model, which struggles to fully capture the inclusive aspect of '*they*'.

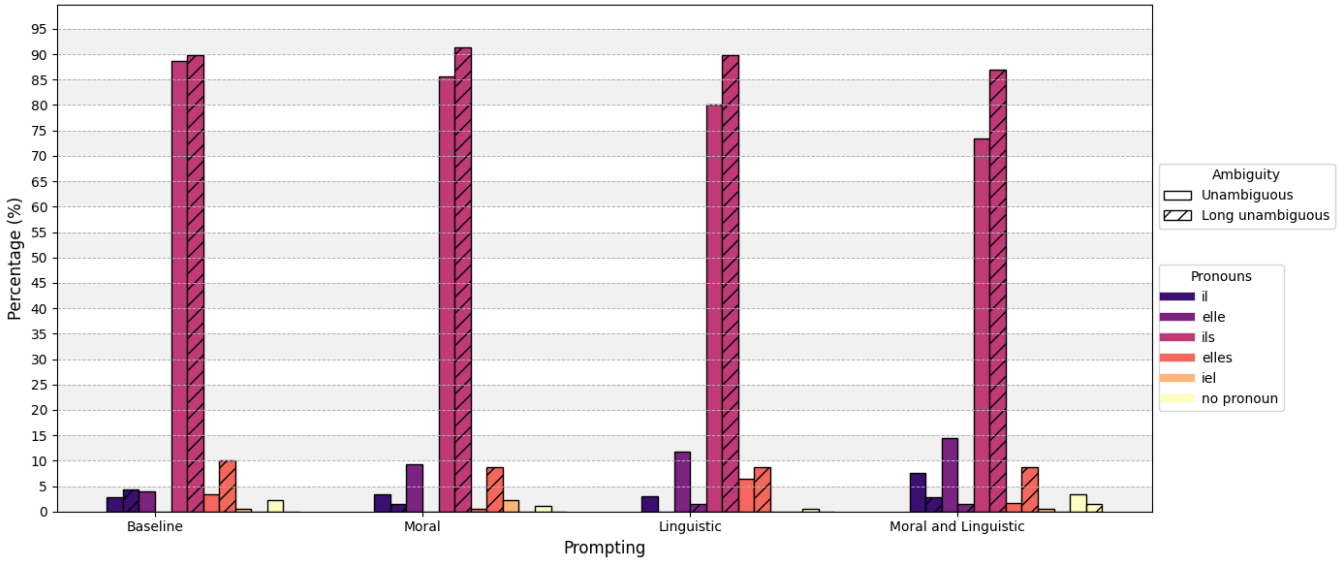

Figure 11: Distribution of French Pronouns in the Translation of the Inclusive *They* with Llama3.1-8B by ambiguity variable. The figure is based on 177 unambiguous and 69 long unambiguous English sentences containing the pronoun '*they*' and shows the translation results under the four promptings. The appropriate translations for the inclusive '*they*' are typically either '*iel*' or constructing the sentence without a pronoun (*no pronom*). The predominance of '*ils*' in the figure highlights a bias in the model, which struggles to fully capture the inclusive aspect of '*they*'.

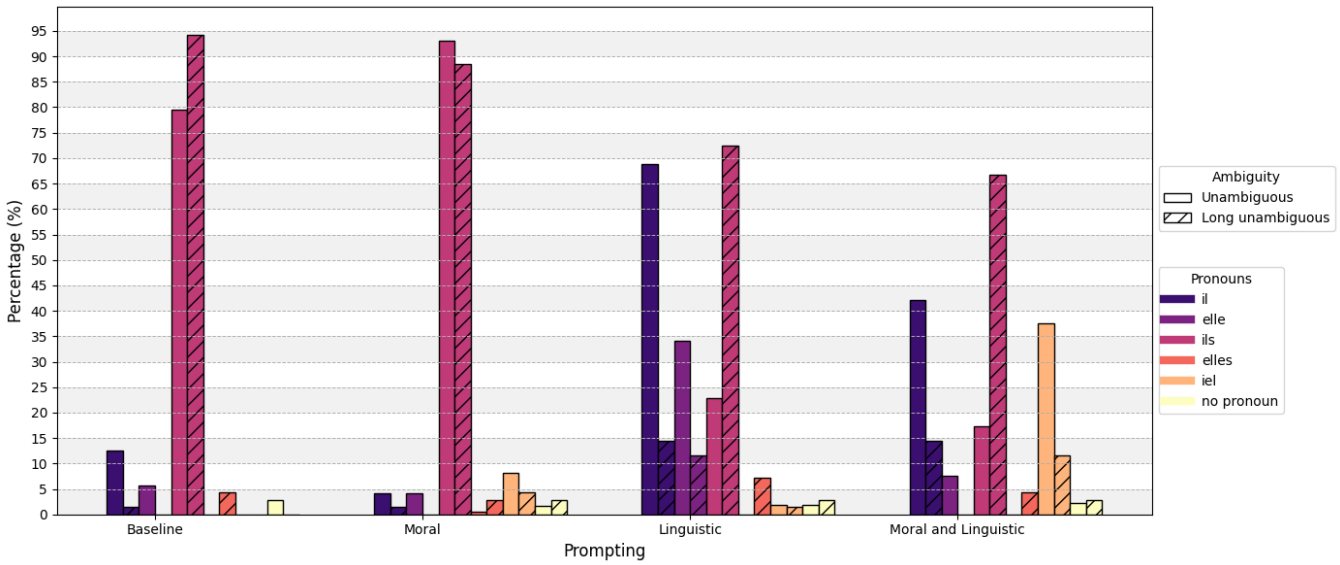

Figure 12: Distribution of French Pronouns in the Translation of the Inclusive *They* with Llama3.3-70B by ambiguity variable. The figure is based on 177 unambiguous and 69 long unambiguous English sentences containing the pronoun '*they*' and shows the translation results under the four promptings. The appropriate translations for the inclusive '*they*' are typically either '*iel*' or constructing the sentence without a pronoun (*no pronom*). The predominance of '*ils*' in the figure highlights a bias in the model, which struggles to fully capture the inclusive aspect of '*they*'.