Rethinking Gender Annotation for Bias Evaluation in Machine Translation: Can LLMs Improve Reliability?

Chiara Manna
Tilburg University
The Netherlands

Argentina Anna Rescigno
University of Pisa
University of Naples "L'Orientale"
Italy

Eva Vanmassenhove
Tilburg University
The Netherlands

{c.manna, e.o.j.vanmassenhove}@tilburguniversity.edu
argentina.rescigno@phd.unipi.it

Abstract

The assessment of gender bias in Machine Translation critically depends on reliable methods for identifying grammatical gender in system outputs. In this paper, we compare the automated WinoMT annotation pipeline, based on word alignment and morphological tagging, with an instruction-tuned LLM (Qwen3-8B) used to annotate grammatical gender in English–Italian translations. Both approaches achieve similar levels of agreement with a human-annotated gold standard, but exhibit distinct systematic weaknesses. The WinoMT pipeline is sensitive to alignment shifts and morphological tagging limitations, often resulting in indeterminate gender labels. The LLM, in contrast, tends to favour binary gender labels even when noun phrases are morphologically gender-invariant. A qualitative analysis of model outputs and reasoning traces further suggests a lack of stable generalization of grammatical gender rules, even when illustrative exemplars are provided for in-context learning. This highlights important methodological limitations in using LLMs as objective annotators for gender evaluation tasks.

1 Introduction

Interest in gender bias in Machine Translation (MT) has grown substantially over the past decade (Savoldi et al., 2025), reflecting broader concerns about systematic biases in Natural Language Processing (NLP) systems (Sun et al., 2019; Costa-jussà, 2019; Blodgett et al., 2020; Stanczak and Augenstein, 2021) as well as specific cross-linguistic challenges inherent to the translation task. Translating from languages in which gender is sparsely marked (e.g., English) into morphologically richer ones (e.g., German, Italian or Hindi) (Stahlberg et al., 2007; Ackerman, 2019; Cao and Daumé III, 2020) often requires making implicit gender information explicit in the target (Vanmassenhove et al., 2018; Vanmassenhove, 2024). This makes gender overt in the target language, thereby exposing the extent to which model outputs reflect statistical patterns (and potential biases) in the training data. As illustrated in Figure 1, even when gender information is available in the context, MT systems tend to default to sta-

INPUT

The cook prepared a soup for the **housekeeper** because **he** helped clean the room.

OUTPUT

Il cuoco ha preparato una zuppa per **la governante** perché ha aiutato a pulire la stanza.

INPUT

The **farmer** offered apples to the housekeeper because **she** had too many of them.

OUTPUT

Il contadino offrì mele alla governante perché ne aveva troppe.

Figure 1: EN→IT WinoMT translation examples defaulting to *la governante* (fem.) and *il contadino* (masc.). Generated by ChatGPT on September 29th, 2025.

tistically dominant gender realizations, rendering *housekeeper* as *la governante* (fem.) and *farmer* as *il contadino* (masc.) in the Italian target.

At the same time, the field of MT has undergone a rapid paradigm shift, transitioning from rule-based systems to statistical and neural MT, and more recently to Large Language Models (LLMs). These developments have significantly improved translation quality and established the new state-of-the-art in MT (Kocmi et al., 2023; Zhu et al., 2024; Kocmi et al., 2024; Deutsch et al., 2025). Nonetheless, the increasing complexity of modern architectures has raised concerns for AI governance regarding their opacity and biases.

While a substantial body of work has focused on analysing MT outputs (Rescigno et al., 2020; Ramesh et al., 2021; Rescigno and Monti, 2023) and developing bias mitigation strategies (Vanmassenhove et al., 2018; Moryossef et al., 2019; Habash et al., 2019; Hirasawa and Komachi, 2019; Font and Costa-jussà, 2019; Zmigrod et al., 2019; Saunders and Byrne, 2020; Vanmassenhove et al., 2021; Sun et al., 2021; Piergentili et al., 2023), significant effort has also been devoted to the development of evaluation frameworks (Stanovsky et al., 2019; Bentivogli et al., 2020; Levy et al., 2021; Manna et al., 2026) that enable a systematic assessment of gender bias and gender-related errors. These frameworks generally compare translation quality across gender conditions, analyse the distribution of gendered outputs in ambiguous contexts, or evaluate whether systems correctly disambiguate grammatical gender in unambiguous contexts.

Evaluation therefore depends not only on the quality of the MT system itself, but also on the reliability of the methods used to identify the realized gender in the output. Many widely adopted benchmarks rely on automatic pipelines, typically involving techniques such as reference matching, word alignment, and morphological analysis. In this context, WinoMT (Stanovsky et al., 2019) has become one of the most widely adopted benchmarks for gender bias evaluation in MT, due to the simplicity and scalability of its integrated pipeline. By combining word alignment (e.g., *fast_align* (Dyer et al., 2013)) with morphological analysis tools (e.g., *spaCy* (Honnibal et al., 2020)), it enables (multilingual) gender annotation without requiring manually constructed references. However, because it relies on two intermediate processing steps, it inevitably introduces a certain degree of noise (alignment or morphological tagging errors) that can affect the automatic assessment and thus the validity of the resulting gender bias measurements. As an alternative, LLMs have demonstrated great potential as high-level annotators for a range of linguistic tasks, including the evaluation of gender-neutral rewriting in MT (Piergentili et al., 2025). Nonetheless, their employment strictly requires rigorous validation against human-annotated gold standards to ensure methodological integrity and validity (Pangakis et al., 2023; Baumann et al., 2025).

To assess the reliability of automatic gender annotation for gender bias evaluation in MT, we compare these two alternative approaches — the WinoMT evaluation pipeline and a state-of-the-art instruction-tuned LLM — against human annotations as gold standard. Specifically, we address the following research questions (RQs).

> **[RQ1]** To what extent do automatic gender annotations produced by the WinoMT pipeline agree with human annotations?

> **[RQ2]** To what extent do automatic gender annotations produced by an instruction-tuned LLM agree with human annotations, and how do they compare with the WinoMT pipeline?

2 Related Work

2.1 Gender Bias Evaluation in MT

Since 2019, a growing number of evaluation frameworks have been proposed to systematically assess gender bias in MT, typically using controlled or contrastive settings to analyse how systems perform at gender disambiguation and whether they default to statistically dominant, often masculine, forms.

Early work mostly relied on synthetic template-based sentences. Among these, Cho et al. (2019) and Prates et al. (2020) investigated the effect of translating from genderless languages (e.g., Hungarian) or languages with gender-neutral pronominal systems (e.g., Korean) into English using minimal template-based sentences of the form *[pronoun] is a [profession]*. In this setup, the realised gender was inferred by manually or automatically

detecting gendered pronouns in the output (e.g., *he* vs. *she*). Importantly, these templates provide no explicit contextual cues for gender at the source level, meaning that any gender assignment in the translation reflects how systems resolve inherently ambiguous inputs.

Subsequent work extended occupation-based templates to grammatical gender languages. Escudé Font et al. (2019), for instance, adopted more complex templates in which the target gender was recovered from morphologically inflected noun phrases (e.g., *amigo* vs. *amiga* in Spanish). Across these approaches, gender annotation relied on lexical or morphological matching procedures, often requiring language-specific rules to detect gender-marked forms in the translated output. Similarly, Stanovsky et al. (2019) introduced WinoMT, a template-based benchmark designed to evaluate gender bias when translating from English into eight grammatical gender languages, including Italian. Compared to earlier template-based approaches, WinoMT introduced more complex (albeit still template-based) sentence structures containing two human entities and a gendered pronoun to disambiguate the gender of the correct referent, as illustrated in Figure 1. In contrast to some of the earlier work (e.g., Cho et al. (2019)), these contextual cues should enable evaluation of whether models follow contextual evidence or still default to stereotypical associations. Moreover, it introduced a unified gender annotation pipeline based on word alignment (e.g., *fast_align* (Dyer et al., 2013)) and automatic morphological tagging (e.g., *spaCy* (Honnibal et al., 2020)) to locate the target entity in the translation output and extract its gender. By removing the need for manually constructed references and language-specific matching strategies, this design enabled, for the first time, scalable multilingual evaluation, contributing to it becoming one of the most widely adopted benchmarks as well as a foundation for numerous extensions (Troles and Schmid, 2021; Levy et al., 2021; Kappl, 2025; Manna et al., 2025; Manna et al., 2026).

Alongside synthetic test sets, several studies have opted for naturally occurring data for the assessment of gender bias in MT (Vanmassenhove et al., 2018; Elaraby and Monz, 2018; Moryossef et al., 2019; Habash et al., 2019; Alhafni et al., 2022; Bentivogli et al., 2020; Vanmassenhove and Monti, 2021). In this context, MuST-SHE (Bentivogli et al., 2020) represents the first structured multilingual benchmark (EN→{IT, ES, FR, DE}), derived from TED talks data. Originally designed to analyse translation quality across speaker and/or referent gender in both ambiguous and unambiguous contexts, it was later adapted to explicitly measure gender accuracy by matching system outputs against predefined gender realizations. Nonetheless, as most of such frameworks, it remains dependent on manually curated references and language-specific matching strategies.

More recently, the WinoMT paradigm has been extended to naturally occurring data through the BUG corpus (Levy et al., 2021), by automatically extracting English sentences linking role nouns with gendered pronouns from sources such as Wikipedia and PubMed. Because BUG preserves the structural properties underlying WinoMT, gender annotation can be performed using the same automatic pipeline across multiple target languages, combining linguistic realism with cross-lingual scalability. As a result, the WinoMT annotation protocol has effectively become a predominant approach for gender bias evaluation across both synthetic and naturally occurring benchmarks. However, the reliability of the underlying pipeline has received limited attention.

Since gender annotation depends on two intermediate automatic processing steps, errors may arise at different stages of the pipeline, inevitably introducing some degree of noise into the evaluation. In addition to alignment errors or morphological tagging failures that may mischaracterise grammatical gender, there are cases in which the pipeline fails to identify any grammatical gender altogether. Such instances are assigned an `UNKNOWN` label and treated as errors under the standard WinoMT evaluation protocol (Stanovsky et al., 2019). As reported by Manna et al. (2026), these cases occur with non-negligible frequency and may distort gender bias metrics as much as incorrect gender labels, despite not necessarily reflecting genuine translation errors.

2.2 LLMs as High-level Annotators

Ultimately, the evaluation of gender bias in MT depends on the quality and limitations of the (automatic) annotation procedure used to identify the realized gender in translated outputs. In this context, recent work has explored the use of LLMs for data annotation in NLP (Pangakis et al., 2023;

Tan et al., 2024), automating complicated labelling processes across various data types (Tan et al., 2024).

While manual annotation is costly, slow, and prone to human fatigue and interpretative shifts (Pangakis et al., 2023; Nasution and Onan, 2024), LLMs might offer a more cost-effective alternative capable of processing complex data using only natural language instructions (Törnberg, 2024). Moreover, they have been shown to rival or even outperform traditional crowd-sourced workers in both speed and inter-annotator agreement in many text-annotation tasks (Gilardi et al., 2023; Pangakis et al., 2023; Nasution and Onan, 2024; Törnberg, 2024). However, their performance still falls short of human experts (as opposed to crowd-sourced workers) in domain-specific settings, indicating that LLMs should not serve as substitutes for expert annotators (Tseng et al., 2025).

Beyond simple text categorization, there is growing interest for the "LLM-as-a-judge" framework to automatically evaluate MT and Natural Language Generation (NLG) (Gao et al., 2025; Li et al., 2025), where LLM evaluators can help overcome the limitations of traditional metrics that rely strictly on n-gram overlap or statistical word alignments (Gao et al., 2025). In the context of gender in translation, recent studies show that LLMs can serve as effective evaluators, particularly when reference translations are available, and that their performance benefits from advanced prompting strategies such as Chain-of-Thought (CoT) prompting (Qian et al., 2024). For instance, Piergentili et al. (2025) demonstrate that LLMs can be successfully deployed to assess gender-neutral translations, with performance improvements under structured prompting. However, this evidence comes from a relatively constrained setting, in which a single phenomenon is annotated and a sentence-level label is produced. Consequently, it does not directly reflect the ability to perform more fine-grained linguistic annotations.

As such, LLMs cannot be blindly assumed to act as objective annotators (Baumann et al., 2025). Their performance is highly dependent on the conceptual difficulty of the task, the specific dataset, and sensitive to the prompt design and task formulation, which may introduce systematic biases and variability in annotation outcomes (Pangakis et al., 2023; Li et al., 2025). Therefore, the need for rigorous validation of LLM-based annotations against human gold standards has been increasingly emphasized in the literature (Pangakis et al., 2023; Törnberg, 2024; Walshe et al., 2025).

3 Experimental Setup

We compare the WinoMT pipeline with an instruction-tuned LLM acting as high-level annotator for grammatical gender in English-to-Italian (EN→IT) translation, by evaluating the extent to which each approach reproduces a human-annotated ground truth. While this research is motivated by the social phenomenon of gender bias, our specific task and gold standard strictly evaluate morphosyntactic (grammatical) gender. The following subsections describe the data, models, prompting configurations, and evaluation protocol used to conduct this comparison.

3.1 Data

We base our experiments on the **WinoMT** challenge set (Stanovsky et al., 2019), which builds on the Winogender (Rudinger et al., 2018) and WinoBias schemas (Zhao et al., 2018). The dataset consists of English sentences following a simple template, in which an ungendered profession noun referring to a human entity (the target noun to be translated) is linked via coreference to a gendered pronoun (e.g., *he* or *she*) that provides the gold gender cue (see Figure 1).

The WinoMT consists of three subsets, pro-stereotypical, anti-stereotypical, and a regular/neutral set. In this work, we focus on the regular set, which counts 3,888 sentences balanced across male and female referents (*i.e.*, gender instantiated via pronouns). From this set, we select a balanced sample of 500 input sentences, ensuring that each profession is equally represented across gold gender conditions (*i.e.*, *he* / *she*) and stereotypical alignment (pro- and anti-stereotypical). Because WinoMT does not provide reference translations and is designed to operate on the output of arbitrary MT systems, we generate translations using two models (§3.2), which serve as input for both annotation approaches as well as our human validation.

3.2 Translation Setup

To generate our target sentences, we translate the selected set of 500 English inputs into Italian (**EN→IT**). This translation direction is particularly relevant for gender bias evaluation in MT,

as gender is not marked in the English source but is typically explicitly realized in the Italian target, creating conditions under which biases may emerge. At the same time, the relatively rich and complex Italian morphology, characterized by both overt gender marking and the presence of gender-invariant (epicene) nouns, makes this setting well-suited for evaluating the reliability of automatic gender annotation methods.

We use two decoder-only models:

1. **Llama2**[1] (Touvron et al., 2023): a general-purpose pretrained LLM, known for strong zero-shot translation performance among open-weight LLMs (Xu et al., 2024).
2. **TowerInstruct**[2] (Alves et al., 2024): an instruction-tuned multilingual model optimized for translation tasks, built on top of Llama2 through (MT-specific) continued pretraining and instruction tuning.

For both models, we select the 7B version, which is both compatible with our computational resources and representative of realistic MT deployment scenarios.

Llama2

```
Translate the following text from
English into Italian.
English: {source_sentence}
Italian: {generated_translation}
```

TowerInstruct

```
<|im_start|> user
Translate the following text from
English into Italian.
English: {source_sentence}
Italian: <|im_end|>
<|im_start|> assistant
{generated_translation}
```

Figure 2: Prompt templates used for the translation task. Plain instruction format for Llama2, chat-style for TowerInstruct.

At this stage of the analysis, our objective is not to optimize translation performance through prompt engineering. Therefore, we adopt a zero-shot prompt format by reproducing the template used for TowerInstruct's post-training[3], and provide a plain natural language equivalent for Llama2 (see Fig. 2). More advanced prompting strategies are instead explored in the LLM-based annotation setup.

3.3 WinoMT Automatic Gender Annotation

The WinoMT evaluation framework relies on an automatic gender annotation pipeline designed to locate the target profession noun in the translation and extract its grammatical gender. To this end, the pipeline consists of two intermediate processing steps. First, word alignment is performed using *fast_align* (Dyer et al., 2013), which establishes a mapping between the source sentence and its translation in order to locate the word corresponding to the English profession noun in the target. Second, morphological analysis is applied using *spaCy* (Honnibal et al., 2020) to extract the grammatical gender feature associated with the aligned profession noun, resulting in one of three labels for each sentence: `MALE`, `FEMALE`, or `UNKNOWN`. The latter represents cases in which no gender marking has been automatically identified, for instance due to gender-invariant forms (e.g., *l'assistente*, *l'autista* in Italian), but also alignment errors, morphological tagging failures, or translation variability (e.g., incoherent output or omission of the target entity). For consistency, these labels are mapped to the `MASCULINE`, `FEMININE`, and `UNKNOWN` scheme used throughout our evaluation.

3.4 LLM-based Gender Annotation

As an alternative automatic approach, we explore the use of an instruction-tuned LLM as high-level annotator for grammatical gender. The model is tasked with reproducing the same annotation procedure as the WinoMT pipeline, namely (i) locating the translated referent corresponding to the English profession noun and (ii) identifying the grammatical gender expressed in the corresponding noun phrase.

For this task, we opt for **Qwen3-8B** (Yang et al., 2025) for several reasons. First, we aim to maintain an open and reproducible gender annotation setup that could potentially serve as a scalable alternative to the WinoMT framework, making open-source models preferable to proprietary alternatives. Second, Qwen3 belongs to the family of reasoning-oriented LLMs, which are specifically designed to perform structured, multi-step inference (Besta et al., 2025). We therefore consider it particularly suitable for our designed gender an-

[1] huggingface.co/meta-llama/Llama-2-7b-hf
[2] huggingface.co/Unbabel/TowerInstruct-7B-v0.2
[3] huggingface.co/datasets/Unbabel/TowerBlocks-v0.2

notation task, which is explicitly decomposed into two steps, as previously defined. This choice is further supported by prior work reporting strong performance of Qwen-family models, in particular Qwen2.5, in similar gender evaluation tasks (Piergentili et al., 2025).

Few-shot exemplars

```
Noun phrases like "il manager", "il
professore", "il direttore"
are MASCULINE.

Noun phrases like "la manager", "la
professoressa", "la direttrice"
are FEMININE.

Noun phrases like "l'assistente",
"l'insegnante", "l'auditor"
are UNKNOWN.
```

Figure 3: Few-shot exemplars used in the LLM-based annotation prompt to illustrate masculine, feminine, and unknown (gender-invariant) noun phrases. For each sentence, three exemplars per label are dynamically filtered so that the profession under annotation does not appear among them.

Our setup is closely related to the one proposed by Piergentili et al. (2025), in which LLMs are employed to assess whether translated referents are correctly rendered in a gender-neutral form. Although they explore different prompting approaches, either producing direct sentence-level assessments or eliciting intermediate phrase-level annotations to be aggregated into a sentence-level judgment, their task implicitly depends on identifying relevant referential expressions and their gender realization, aligning closely with our annotation task. Building on their framework, we adopt a comparable prompting strategy, in both **zero-shot** and **few-shot** configurations, in which the model receives the English source sentence, its Italian translation, and the English profession noun corresponding to the referent of interest. While the zero-shot setup relies exclusively on natural language task instructions, the few-shot configuration augments these with three illustrative examples per label, namely Italian noun phrases representing masculine, feminine, and gender-invariant forms (e.g., *il professore*, *la direttrice*, *l'assistente*), to elicit in-context learning (see Fig. 3). To avoid lexical leakage, exemplars are selected dynamically so that the profession noun under annotation is never included among these.

In both configurations, the model is instructed to identify the translated noun phrase and assign a gender label based exclusively on morphosyntactic cues internal to the target noun phrase (e.g., determiner and noun morphology), to replicate the WinoMT protocol. We further constrain model outputs to a fixed sentence-level label format (*i.e.*, `MASCULINE`, `FEMININE`, or `UNKNOWN`). Both prompting configurations are applied to the same set of 1,000 translated sentences (two sets of 500 sentences) and the full (system) prompts are provided in the Appendix.

3.5 Human Gold Annotation

To evaluate the reliability of both the WinoMT automatic pipeline and the LLM-based annotations, we use a human annotation as our gold standard. An expert linguist fluent in the Italian language manually annotated the target sentences. The annotator was instructed to label the grammatical gender of the target profession noun, by looking at the determiner and the noun itself. Therefore, the task is strictly formal and deterministic, leaving no space for ambiguity or subjectivity.

The same label organisation is adopted throughout the analysis by assigning `MASCULINE` or `FEMININE` according to the grammatical gender expressed by the profession noun, and `UNKNOWN` whenever gender cannot be reliably inferred, even when considering the associated determiner. Since this classification is based on the Italian language morphosyntactic rules of agreement, it is free from personal interpretation and ensures objective and consistent labelling across the datasets. The creation of this objective gold standard allowed us to compute the **agreement with the human annotation** as a direct measure of the reliability of each automated approach.

Dataset	MASC.	FEM.	UNKNOWN
Llama2	0.70	0.18	0.12
TowerInstruct	0.56	0.33	0.11

Table 1: Distribution of (human) gold annotation labels across the translated outputs generated by each MT system. Proportions are computed over the 500 annotated sentences.

Although our sample is balanced with respect to source-side gender cues (Section 3.1), our MT outputs are not. As reported in Table 1, we can observe that both Llama2 and TowerInstruct exhibit a strong tendency to default to masculine realizations, resulting in substantially skewed gender distributions in our target sentences, which reflect realistic MT deployment scenarios. There-

Dataset	Method	Accuracy	Macro F1	Class-specific F1		
				Masculine	Feminine	Unknown
Llama2	WinoMT	0.92	**0.85**	0.95	0.94	**0.67**
	LLM zero-shot	0.90	0.75	0.96	0.90	0.40
	LLM few-shot	0.92	0.79	0.96	0.92	0.50
TowerInstruct	WinoMT	0.88	**0.82**	0.93	0.91	**0.63**
	LLM zero-shot	0.91	0.78	0.95	0.94	0.45
	LLM few-shot	0.92	**0.82**	0.96	0.95	0.57

Table 2: Agreement with human gold annotations across automatic annotation methods. We report overall accuracy, macro F1, and per-class F1 scores on the Llama2- and TowerInstruct-generated translation sets.

fore, agreement scores are computed not only in terms of overall accuracy, but also using macro and per-class F1 scores, in order to better capture label-specific annotation behaviour.

4 Results

To establish the baseline for our evaluation (RQ1), we first measure the agreement between the **WinoMT pipeline** and our manually annotated gold standard. As reported in Table 2, the integrated pipeline achieves strong overall accuracy scores, reaching 0.92 on the Llama2-translated set and 0.88 on the TowerInstruct-translated one. However, the F1-based evaluation reveals important label-specific asymmetries. In particular, the class-specific F1 breakdown indicates uneven performance across labels, with substantially weaker scores for the UNKNOWN category.

To answer RQ2, we then evaluate the performance of **Qwen3-8B** acting as grammatical gender annotator in both a zero-shot and a few-shot setting. In the **zero-shot** configuration, the LLM's performance is virtually on par with the WinoMT pipeline in terms of overall accuracy. However, the F1-based evaluation reveals a strong tendency to force binary gender labels onto morphologically gender-invariant nouns. While performance remains high for MASCULINE and FEMININE instances (with class-specific F1 scores between 0.90 and 0.96), it drops substantially for the UNKNOWN category (with scores between 0.40 and 0.45). Accordingly, incorrect gender annotations most frequently involve cases labelled as UNKNOWN in the human gold standard (Tab 4).

To try and mitigate this by clarifying the expected decision boundary between morphologically marked and invariant forms, we evaluate the performance of the **few-shot** prompting strategy including specific annotations exemplars for each label. While overall accuracy remains largely comparable to the zero-shot setting, the few-shot configuration leads to a modest increase in the (macro) F1 score, primarily driven by improved performance on the UNKNOWN category.

5 Analysis and Discussion

While overall accuracy scores suggest relatively comparable performance across methods, a more fine-grained evaluation exposes qualitative differences in annotation behaviour.

(a) Llama2

Predicted label	WinoMT	LLM zero-shot	LLM few-shot
Masculine	0.46	**0.54**	**0.57**
Feminine	0.12	0.38	0.36
Unknown	**0.41**	0.08	0.07
Total errors (n)	41	48	42

(b) TowerInstruct

Predicted label	WinoMT	LLM zero-shot	LLM few-shot
Masculine	0.22	**0.53**	**0.53**
Feminine	0.19	0.30	0.25
Unknown	**0.59**	0.17	0.23
Total errors (n)	59	47	40

Table 3: Distribution of errors by predicted label across annotation methods for the Llama2- and TowerInstruct-generated datasets. Values are proportions computed over the total number of errors per method.

A closer analysis of the **WinoMT pipeline** reveals an important limitation stemming from its reliance on word alignment and automatic morphological tagging. In our formulation, the UNKNOWN label is intended to capture cases in which gender cannot be determined (e.g., for morphologically gender-invariant, epicene forms). In practice, however, the pipeline also assigns UNKNOWN when gender cannot be recovered due to alignment shifts, translation variability, or tagging errors. As a result, the UNKNOWN category conflates genuine linguistic indeterminacy with pipeline failures. This behaviour is reflected in the low class-

specific F1 scores observed for the UNKNOWN category (Tab. 2), as well as the distribution of *errors by predicted label* reported in Table 3. Most incorrect annotations produced by the WinoMT pipeline indeed fall under the UNKNOWN label, indicating a systematic tendency to overproduce indeterminate annotations.

Although this lies beyond the scope of the present work, these observations have important implications for the interpretation of Wino-MT based evaluations. Under the WinoMT protocol, UNKNOWN instances are treated as "incorrectly gendered" and therefore penalized in the computation of gender bias metrics. As already discussed by Manna et. al (2026), such a design choice may distort the resulting assessment, especially considering that neither morphologically gender-invariant forms nor annotation errors necessarily reflect bias in the MT system itself.

(a) Llama2

Gold label	LLM zero-shot	LLM few-shot
Masculine	0.04	0.05
Feminine	0.04	0.02
Unknown	**0.92**	**0.93**
Total Errors (n)	48	42

(b) TowerInstruct

Gold label	LLM zero-shot	LLM few-shot
Masculine	0.04	0.08
Feminine	0.15	0.15
Unknown	**0.81**	**0.78**
Total Errors (n)	47	40

Table 4: Distribution of LLM annotation errors by gold label for the Llama2- and TowerInstruct-generated sets under zero-shot and few-shot prompting. Values are proportions computed over the total number of errors per method.

When prompting **Qwen3-8B** for the same annotation task, we observe accuracy scores comparable to those obtained with the WinoMT pipeline. However, the F1-based evaluation reveals a different weakness. In the zero-shot configuration, the LLM exhibits a strong bias toward enforcing a binary choice, assigning either MASCULINE or FEMININE even when the (Italian) noun phrase is morphologically gender-invariant. This behaviour is reflected in the markedly lower class-specific F1 scores obtained for the UNKNOWN category, as well as in the distribution of *errors by predicted label* (Tab. 3), which shows that incorrect annotations are predominantly assigned to binary categories, with a strong skew toward MASCULINE (and to a lesser extent FEMININE). At the same time, the distribution of *errors by gold label* (Tab. 4) indicates that the vast majority of these errors involve instances labelled as UNKNOWN in the human annotation. Taken together, these observations suggest that the LLM struggles to recognize genuine linguistic indeterminacy, instead defaulting to a binary annotation.

While overall accuracy remains largely unchanged, augmenting the prompt with illustrative examples leads to a slight improvement in both macro and class-specific F1 scores for the UNKNOWN category (Tab. 2). Nonetheless, this effect is inconsistent: the few-shot configuration both corrects and introduces errors. While some incorrect binary annotations are revised to UNKNOWN, some cases that were (previously) correctly identified as UNKNOWN are reassigned to a binary category.

A qualitative inspection of the generated reasoning (Fig. 4), further highlights this inconsistency, revealing that correct annotations do not

Case 1: binary → UNKNOWN

*Il supervisore criticò **l'assistente** perché ha commesso degli errori.*
lit. *"The supervisor criticised the attendant because they made some mistakes."*

zero-shot FEMININE
"l' corresponds to la, therefore feminine."

few-shot UNKNOWN
"assistente can be masculine or feminine … similar to l'insegnante."

Case 2: UNKNOWN → binary

*Il meccanico chiese **l'assistente** perché era l'unico che poteva aiutarlo.*
lit. *"The mechanic asked for the attendant because they were the only one who could help."*

zero-shot UNKNOWN
"assistente can be either masculine or feminine."

few-shot FEMININE
"assistente ends in -e, typically feminine."

Figure 4: Examples of annotation flips and corresponding reasoning (excerpts) between zero-shot and few-shot prompting for the same morphologically (gender-)invariant noun phrase (*l'assistente*). Highlighted labels indicate agreement (green) or disagreement (orange) with human gold annotation.

directly derive from a stable understanding of (Italian) morphological gender marking. For instance, the same morphologically gender-invariant noun phrase (*l'assistente*) receives different labels across prompting conditions, often accompanied by inconsistent or contradictory explanations. In some cases, the model correctly identifies the noun as gender-invariant, while in others it incorrectly infers a binary label based on simplified or inaccurate heuristics (e.g., associating the ending *-e* with feminine gender).

More generally, the model often produces linguistically inaccurate or partially inconsistent explanations that may nevertheless lead to the correct label by relying on surface-level morphosyntactic patterns rather than a stable generalization of grammatical gender rules, even when illustrative exemplars are provided.

6 Conclusion

The study examines the reliability of automatic gender annotation methods for assessing bias in machine translation, by comparing the widely adopted automated WinoMT pipeline with an instruction-tuned LLM, Qwen3-8B, prompted to reproduce the same multi-step task.

While both approaches achieve comparable accuracy scores against human gold annotations, they exhibit distinct systematic weaknesses. The WinoMT pipeline tends to overassign the `UNKNOWN` label due to alignment errors or morphological tagging limitations on complex morphological constructions. In contrast, the LLM tends to favour binary gender annotations (`MASCULINE` or `FEMININE`), even when faced with ambiguous or epicene Italian nouns (e.g. *insegnante*, *assistente*). We further observe that the introduction of specific examples in the prompt leads to only marginal changes in overall agreement with human annotations. A thorough examination of the reasoning process generated by the model further reveals that correct annotations are not consistently supported by a genuine understanding of Italian morphological rules.

Overall, while LLM-based (gender) annotation proves to be a competitive alternative to the WinoMT pipeline, it cannot be considered infallible or a substitute for human annotation. Its dependence on prompt formulation and the lack of solid linguistic reasoning confirm the necessity of a constant validation against human gold standard annotations, to preserve the integrity and reliability of gender bias assessment.

7 Limitations

We acknowledge that our findings are limited to Italian, a grammatical gender language with relatively rich and transparent gender morphology. Both investigated approaches may behave differently for languages in which gender is more sparsely marked, as well as in low-resource settings, where less consistent agreement patterns and limited representation in the training data may reduce the reliability and robustness of automatic gender annotation. Similarly, the LLM-based annotation setup is evaluated using a single instruction-tuned LLM, and different models may exhibit different annotation behaviours and sensitivities to prompting strategies.

A further limitation concerns the treatment of gender-neutral language. While in our analysis the `UNKNOWN` label is interpreted as reflecting gender-invariant forms (e.g., epicene nouns), gender-neutrality in Italian is often achieved through reformulation strategies such as pluralization, impersonal constructions, or other structural changes (Sulis and Gheno, 2022; Piergentili et al., 2023). These strategies may not be captured by morphology-based annotation alone, and alignment-based methods may struggle to map complex reformulations to the source referent.

Finally, it is important to note that the annotation task considered in this work remains relatively constrained. It involves a single phenomenon (gender marking), which is annotated based on a limited set of labels, and can therefore be categorized as a high-level classification task. As such, our findings should not be interpreted as evidence of LLMs' ability (or inability) to perform more fine-grained linguistic annotations. Rather, they reflect the behaviour of such systems within a simplified and controlled evaluation setting.

8 Acknowledgements

The second author acknowledges funding from the National PhD programme in Artificial Intelligence, partnered by the University of Pisa and the University of Naples "L'Orientale", through the doctoral grant 39-411-24-DOT23A27WJ-7219 established by Ex DM 318, of type 4.1, co-financed by the National Recovery and Resilience Plan.

References

Ackerman, Lauren. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1).

Alhafni, Bashar, Nizar Habash, and Houda Bouamor. 2022. The Arabic parallel gender corpus 2.0: Extensions and analyses. In Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France, June. European Language Resources Association.

Alves, Duarte M, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Baumann, Joachim, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation. *arXiv preprint arXiv:2509.08825*.

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

Besta, Maciej, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, Tomasz Sternal, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Łukasz Flis, Hannes Eberhard, Zixuan Chen, Hubert Niewiadomski, and Torsten Hoefler. 2025. Reasoning language models: A blueprint.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July. Association for Computational Linguistics.

Cao, Yang Trista and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595. Association for Computational Linguistics, July.

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. Association for Computational Linguistics.

Costa-jussà, Marta R. 2019. An analysis of Gender Bias studies in Natural Language Processing. *Nature Machine Intelligence*, 1(11):495–496.

Deutsch, Daniel, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Dyer, Chris, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 644–648.

Elaraby, Mohamed and Christof Monz. 2018. Gender-aware neural machine translation. In *Proceedings of the Second International Conference on Natural Language and Speech Processing (ICNLSP)*. IEEE.

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.

Gao, Mingqi, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Habash, Nizar, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165.

Hirasawa, Tosho and Mamoru Komachi. 2019. Debiasing word embeddings improves multimodal machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 32–42.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Kappl, Michelle. 2025. Are all spanish doctors male? evaluating gender bias in german machine translation. *arXiv preprint arXiv:2502.19104*.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December. Association for Computational Linguistics.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA, November. Association for Computational Linguistics.

Levy, Shahar, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Li, Haitao, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. 2025. CalibraEval: Calibrating prediction distribution to mitigate selection bias in LLMs-as-judges. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16537–16552, Vienna, Austria, July. Association for Computational Linguistics.

Manna, Chiara, Afra Alishahi, Frédéric Blain, and Eva Vanmassenhove. 2025. Are we paying attention to her? investigating gender disambiguation and attention in 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 1–16, Geneva, Switzerland, June. European Association for Machine Translation.

Manna, Chiara, Hosein Mohebbi, Afra Alishahi, Frédéric Blain, and Eva Vanmassenhove. 2026. Gender disambiguation in machine translation: Diagnostic evaluation in decoder-only architectures. *arXiv preprint arXiv:2603.17952*.

Moryossef, Amit, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54.

Nasution, Arbi Haza and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900.

Pangakis, Nicholas, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*.

Piergentili, Andrea, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In Vanmassenhove, Eva, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors, *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland, June. European Association for Machine Translation.

Piergentili, Andrea, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025. An llm-as-a-judge approach for scalable gender-neutral translation evaluation.

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

Qian, Shenbin, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orăsan, Tharindu Ranasinghe, and Frédéric Blain. 2024. What do large language models need for machine translation evaluation?

Ramesh, Krithika, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in hindi-english machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23.

Rescigno, Argentina Anna and Johanna Monti. 2023. Gender bias in machine translation: a statistical evaluation of google translate and deepl for english, italian and german. *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*.

Rescigno, Argentina Anna, Eva Vanmassenhove, Johanna Monti, and Andy Way. 2020. A case study of natural gender phenomena in translation a comparison of google translate, bing microsoft translator

and deepl for english to italian, french and spanish. *Computational Linguistics CLiC-it 2020*, page 359.

Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

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.

Savoldi, Beatrice, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6).

Stahlberg, Dagmar, F. Braun, L. Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social Communication*, pages 163–187, 01.

Stanczak, Karolina and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

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

Sulis, Gigliola and Vera Gheno. 2022. The debate on language and gender in italy, from the visibility of women to inclusive language (1980s–2020s). *The Italianist*, 42(1):153–183.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

Sun, Tony, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral english. *arXiv preprint arXiv:2102.06788*.

Tan, Zhen, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Troles, Jonas-Dario and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online, November. Association for Computational Linguistics.

Tseng, Yu-Min, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2025. Evaluating large language models as expert annotators. *arXiv preprint arXiv:2508.07827*.

Törnberg, Petter. 2024. Best practices for text annotation with large language models.

Vanmassenhove, Eva and Johanna Monti. 2021. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In Costa-jussà, Marta R., Hila Gonen, Christian Hardmeier, and Kellie Webster, editors, *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online, August. Association for Computational Linguistics.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

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

Vanmassenhove, Eva. 2024. Gender bias in machine translation and the era of large language models. In *Gendered Technology in Translation and Interpreting*, pages 225–252. Routledge.

Walshe, Thomas, Sae Young Moon, Chunyang Xiao, Yawwani Gunawardana, and Fran Silavong. 2025. Automatic labelling with open-source llms using dynamic label schema integration.

Xu, Haoran, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *International Conference on Learning Representations (ICLR) 2024*.

Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.

Zmigrod, Ran, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

A Appendix

LLM-based Annotation Prompts

We report below the full system prompts used for the zero-shot and few-shot annotation configurations. In the few-shot configuration, three exemplars per label are dynamically selected such that the profession under annotation does not appear among them.

Zero-shot System Prompt

```
You are a language expert specializing in grammatical gender in translations from
English into Italian.

Your task is to identify the grammatical gender expressed in the Italian
translation of a given English profession.

You will receive:
- An English source sentence
- Its Italian translation
- The English profession noun

Instructions:
1. Locate the Italian translation of the specified profession.
2. Identify the noun phrase (determiner + profession noun).
3. Determine the grammatical gender expressed in that noun phrase.

Use ONLY:
- The determiner
- The morphological form of the noun

If the noun itself does not present any gender marking but the determiner does,
report the gender of the latter.

Do NOT use:
- Pronouns
- Adjectives outside the noun phrase
- Verbal agreement
- Context
- English gender cues
- World knowledge

Assign one of the following labels:
- MASCULINE: if the determiner and noun clearly mark masculine gender.
- FEMININE: if the determiner and noun clearly mark feminine gender.
- UNKNOWN: if the noun phrase is gender-invariant, lacks gender marking, or
  shows conflicting signals.

If uncertain, output UNKNOWN.

Output format:
Label: <MASCULINE | FEMININE | UNKNOWN>

Return only this line. Do not provide explanations.
```

Few-shot System Prompt

```
You are a language expert specializing in grammatical gender in translations from
English into Italian.

Your task is to identify the grammatical gender expressed in the Italian
translation of a given English profession.

You will receive:
- An English source sentence
- Its Italian translation
- The English profession noun

Instructions:
1. Locate the Italian translation of the specified profession.
2. Identify the noun phrase (determiner + profession noun).
3. Determine the grammatical gender expressed in that noun phrase.

Use ONLY:
- The determiner
- The morphological form of the noun

If the noun itself does not present any gender marking but the determiner does,
report the gender of the latter.

Do NOT use:
- Pronouns
- Adjectives outside the noun phrase
- Verbal agreement
- Context
- English gender cues
- World knowledge

Assign one of the following labels:
- MASCULINE: if the determiner and noun clearly mark masculine gender.
- FEMININE: if the determiner and noun clearly mark feminine gender.
- UNKNOWN: if the noun phrase is gender-invariant, lacks gender marking, or
  shows conflicting signals.

If uncertain, output UNKNOWN.

Noun phrases like "il manager", "il professore", "il direttore" are MASCULINE.
Noun phrases like "la manager", "la professoressa", "la direttrice" are FEMININE.
Noun phrases like "l'assistente", "l'insegnante", "l'auditor" are UNKNOWN.

Output format:
Label: <MASCULINE | FEMININE | UNKNOWN>

Return only this line. Do not provide explanations.
```