Can Explicit Gender Information Improve Zero-Shot Machine Translation?

Van-Hien Tran* **Huy Hien Vu*** **Hideki Tanaka** **Masao Utiyama**
National Institute of Information and Communications Technology (NICT)
{tran.vanhien, vuhuyhien, hideki.tanaka, mutiyama}@nict.go.jp

Abstract

Large language models (LLMs) have demonstrated strong zero-shot machine translation (MT) performance but often exhibit gender bias that is present in their training data, especially when translating into grammatically gendered languages. In this paper, we investigate whether explicitly providing gender information can mitigate this issue and improve translation quality. We propose a two-step approach: (1) inferring entity gender from context, and (2) incorporating this information into prompts using either **Structured Tagging** or **Natural Language**. Experiments with five LLMs across four language pairs show that explicit gender cues consistently reduce gender errors, with structured tagging yielding the largest gains. Our results highlight prompt-level gender disambiguation as a simple yet effective strategy for more accurate and fair zero-shot MT.

1 Introduction

Large language models (LLMs) have exhibited impressive capabilities in zero-shot machine translation (MT) by leveraging cross-lingual patterns acquired during pretraining (Tran and Utiyama, 2025). However, these models also inherit and propagate societal biases present in their training data, leading to systematic gender mistranslations (Sant et al., 2024). This issue is especially pronounced when translating from languages without grammatical gender into those with gendered grammatical systems (Ghosh and Caliskan, 2023; Tran et al., 2023; Piergentili et al., 2024).

Gender bias in LLM-based MT can be observed when models incorrectly assign gender in translations, even when the source sentence provides sufficient contextual clues to infer the correct gendered form (Vanmassenhove, 2024; Portillo-Palma and Alvarez-Vidal, 2024). For instance, given the sentence, "*The carpenter built the attendant a desk as a gesture of her love.*", an LLM might translate "*carpenter*" into the masculine German form "*der Schreiner*" rather than the correct feminine form "*die Schreinerin*". Such errors highlight a failure to leverage clear syntactic and semantic cues in the source text. To ensure accurate and fair translations, it is essential for LLMs to first resolve gender disambiguation from context before performing translation.

In this work, we investigate whether explicitly incorporating gender information derived from contextual cues during prompting can help LLMs mitigate inherited gender biases when translating into grammatically gendered languages, thereby enhancing overall translation quality. We focus on sentences in which syntactic cues, such as gendered pronouns, unambiguously indicate the gender of an entity, yet may conflict with prevailing societal stereotypes. We hypothesize that making this gender information explicit enables LLMs to rely more heavily on linguistic evidence rather than stereotypical associations, resulting in more accurate and equitable translations.

To evaluate this hypothesis, we propose a two-step prompting framework. In the first step, we leverage LLMs' own capabilities to infer the gender of entities from context alone. In the second step, this inferred gender information is incorporated into the translation prompt to explicitly guide the model. Inspired by the work of Vu et al. (2024); Tran et al. (2025), in which additional information can solve MT tasks in various aspects, we explore two strategies for conveying this information: **Structured Tagging**, which uses formal markers, and **Natural Language**, which embeds gender cues within fluent descriptive text. Extensive experiments across five LLMs and four language pairs show that our explicit gender prompting approach consistently improves translation quality and reduces gender-related errors. Among the two strategies, structured tagging yields the best improve-

*Equal contribution.

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 171–181
August 1, 2025

ments, demonstrating its effectiveness in promoting accurate gender realization and more reliable translations.

2 Related Work

Gender bias has been shown issues to various fields in Natural Language Processing (Blodgett et al., 2020) under different settings and tasks, i.e from foundation model (Dev et al., 2020; Bender et al., 2021; Kaneko et al., 2022) to specific tasks Question Answering (Li et al., 2020; Parrish et al., 2022), Coreference Resolution (Rudinger et al., 2018; Zhao et al., 2018) and others (Sheng et al., 2019; Dev et al., 2020). In the era of large language models (LLMs), the research community has analyzed the impact of gender bias (Kotek et al., 2023; Chen et al., 2025) and proposed several mitigation strategies. These include parameter-based approaches such as fine-tuning (Raza et al., 2024; Zhang et al., 2024), controlled decoding (Liu et al., 2021), and model editing (Cai et al., 2024), as well as prompt-based methods like using specially designed structures, i.e. chain-of-thought prompting, in-context learning, etc. (Schick et al., 2021; Sant et al., 2024; Qiu et al., 2025)

In the field of MT, gender bias has been shown to negatively affect translation quality (Savoldi et al., 2024; Sant et al., 2024; Gete and Etchegoyhen, 2024; Sánchez et al., 2024), often leading to incorrect or stereotypical gender representations in target languages (Li et al., 2020; Farkas and Németh, 2022; Kostikova et al., 2023). To support research in this area, several benchmark datasets and evaluation resources have been developed to facilitate systematic analysis of gender-related translation errors (Currey et al., 2022; Mastromichalakis et al., 2024). In response, various mitigation strategies have been proposed, with a main focus on fine-tuning, balancing genders in dataset, adaptive learning method and prompting (Escudé Font and Costa-jussà, 2019; Costa-jussà and de Jorge, 2020; Saunders and Byrne, 2020; Qiu et al., 2025).

3 Our Approach

This study addresses the challenge of translating source sentences from languages without grammatical gender (e.g., English) into target languages that exhibit grammatical gender distinctions (e.g., German). Specifically, we focus on cases involving gender-unambiguous entities, those whose gender can be reliably inferred from contextual information present within the sentence itself.

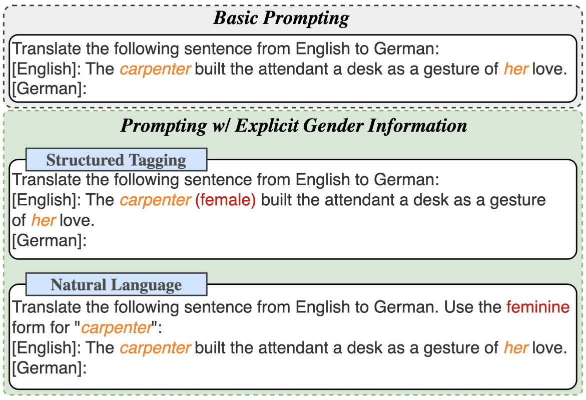

Figure 1: The overall framework.

We define a *gender-unambiguous* entity as one for which the grammatical gender can be deduced unambiguously from intra-sentential cues. For instance, consider the English sentence in Figure 1: "*The carpenter built the attendant a desk as a gesture of her love.*". In this example, the pronoun *her* clearly identifies the referent *carpenter* as female. An accurate and contextually faithful translation into a gendered target language should appropriately reflect this gender information in its lexical and grammatical choices.

We employ a basic prompting template for zero-shot MT. The template is structured as follows: "Translate the following sentence from *src_lang* to *tgt_lang*:\n[*src_lang*]: *A*\n[*tgt_lang*]: " In this format, *src_lang* and *tgt_lang* denote the source and target languages, respectively, while *A* represents the input sentence to be translated. An illustrative example of this prompting strategy is provided in the upper portion of Figure 1.

In contrast to the baseline, our proposed two-step approach integrates explicit gender inference into the translation process. In the first step, we leverage the capabilities of LLMs to infer the gender of a target entity based exclusively on contextual information provided in the source sentence. Let $\mathcal{M}$ denote a decoder-only model, primarily trained on English text. Given $\mathcal{M}$'s deep understanding of English, we leverage it to predict the gender of the entity. We employ the following prompting template $\mathcal{T}$:

"You are an expert in English grammar and linguistic analysis, with no gender bias. Based solely on the information provided in the context, determine the most likely gender (male or female) of the specified entity.

[Context]: The cook prepared a dish for the teacher because he is hungry.
[Conclusion]: In this context, the gender of 'teacher' is male.

[Context]: The mover outran the hairdresser because she was in shape.
[Conclusion]: In this context, the gender of 'mover' is female.

[Context]: The hairdresser befriended the cook because he liked the food.
[Conclusion]: In this context, the gender of 'hairdresser' is male.

[Context]: *A*
[Conclusion]: In this context, the gender of '*B*' is "

Here, *A* refers to the source sentence, and *B* denotes the entity of interest within that sentence. Given the candidate set {`male`, `female`}, our objective is to accurately infer the gender of *B* based on contextual cues in *A*.

For each gender candidate, we combine $\mathcal{T}$ with the gender candidate to create a full statement. This statement is then tokenized into N tokens: $w_1, w_2, \ldots, w_{N_1}, w_{N_1+1}, \ldots, w_N$. The first N_1 tokens come from $\mathcal{T}$, while the rest are from the gender candidate. We calculate the perplexity only over the $(N - N_1)$ tokens of the gender candidate in the full statement:

$$\mathrm{PPL}_{\text{cand}} = \exp\left(-\frac{1}{N-N_1}\sum_{i=N_1+1}^{N}\log P_{\mathcal{M}}(w_i \mid w_1, \ldots, w_{i-1})\right)$$

Here, $P_{\mathcal{M}}(w_i \mid w_1, \ldots, w_{i-1})$ is the probability of token w_i given its preceding context as estimated by the model $\mathcal{M}$. After computing the perplexity scores for both gender candidates associated with the entity *B*, we select the candidate with the lowest perplexity as the predicted gender: $\hat{G} = \arg\min_{j\in\{1,2\}} \mathrm{PPL}(G_j)$.

We incorporate the predicted gender information into the translation prompt, as illustrated in the lower portion of Figure 1, using two distinct formatting strategies: **Structured Tagging** and **Natural Language**. By explicitly including a single, high-confidence gender prediction, we aim to enhance the model's ability to accurately reflect the gender of the target entity during translation.

In the **Structured Tagging** approach, the gender information is appended directly adjacent to the entity using bracket notation (e.g., carpenter (female)). In contrast, the **Natural Language** approach conveys the same information in the form of a natural language instruction, such as: "*Use the feminine form for 'carpenter'*" for female referents, and "*Use the masculine form for 'carpenter'*" for male referents.

It is important to note that, following translation using the **Structured Tagging** method, we apply a post-processing step to remove bracketed gender annotations (e.g., "(female)" or "(male)") from the translated output. This is achieved through a simple heuristic based on dictionary lookup to identify and omit corresponding phrases in the target language, ensuring that the final translation remains natural and fluent.

4 Experiments

4.1 Dataset and Settings

Dataset We use the WinoBias benchmark dataset (Zhao et al., 2018). This dataset contains English sentences, where each sentence contains one entity with a pronoun that refers to it. For our experiments, we selected sentences where the pronoun clearly reflects the gender of the entity (e.g., *him*, *her*, *he*, *she*, ...). For MT task, we evaluate LLMs on translating these sentences into four target languages: German, Italian, Portuguese, and Spanish.

Settings We evaluate our approach on five instruction-tuned LLMs that differ in their pretraining language distributions including Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct, Qwen2.5-72B-Instruct (Qwen et al., 2025).The Llama-family models focus more on English, whereas the Qwen-family models have a more balanced mix of English and Chinese text. For brevity, we refer to the models as Llama 3.2 3B, Llama 3.1 8B, Llama 3.1 72B, Qwen 2.5 7B, and Qwen 2.5 72B throughout this paper. We keep all LLM parameters frozen during the experiments. For text generation, we use non-sampling greedy decoding, a maximum of 256 new tokens, and BF16 precision. Each experiment runs on a machine with eight NVIDIA A100 40GB GPUs.

	En-De↑					En-It↑				
	Base	+T	+N	+WT	+WN	Base	+T	+N	+WT	+WN
Llama 3.2 3B	**80.85**	80.56	80.25	79.88†	79.49†	83.31	83.42	**83.53**	83.03†	82.84†
Llama 3.1 8B	83.23	**83.50***	83.37	82.64†	82.02†	83.75	84.34*	**84.43***	83.61	82.92†
Llama 3.1 70B	84.10	**84.51***	84.49*	83.74†	82.76†	84.41	**84.61**	83.77	83.65†	83.44†
Qwen 2.5 7B	81.51	**81.77**	81.51	80.73†	79.67†	81.86	**82.65***	81.41	81.31†	80.11†
Qwen 2.5 72B	82.77	83.60*	**83.96***	82.68	81.19†	81.97	82.50	**83.16***	81.33†	80.84†
	En-Pt↑					**En-Es↑**				
	Base	+T	+N	+WT	+WN	Base	+T	+N	+WT	+WN
Llama 3.2 3B	82.09	**82.45***	82.04	81.00†	81.58†	83.28	**84.06***	83.69	83.04†	82.57†
Llama 3.1 8B	82.75	**83.57***	83.45*	82.28†	82.07†	85.35	**85.79***	85.51	84.81†	84.29†
Llama 3.1 70B	84.02	84.45*	**84.63***	83.03†	82.46†	85.82	**86.12***	86.10*	85.53†	84.99†
Qwen 2.5 7B	82.98	**83.49***	83.48*	81.94†	81.43†	**84.65**	84.59	83.47	83.12†	82.47†
Qwen 2.5 72B	83.27	**84.57***	84.14*	82.30†	80.45†	85.52	**85.85***	85.73*	84.74†	83.63†

(*) indicates statistical significance at $p < 0.05$ when comparing the *+T* and *+N* systems to the *Base* system.
(†) indicates statistical significance at $p < 0.05$ when comparing the *Base* system to the *+WT* and *+WN* systems.

Table 1: The results of main experiments for English-German (En-De), English-Italian (En-It), English-Portuguese (En-Pt) and English-Spanish (En-Es) datasets. The best performance per metric are in bold text.

To examine the impact of incorporating explicit gender information, we compare the baseline model (Base) with our proposed methods using **Structured Tagging** (+T) and **Natural Language** (+N), as presented in Table 1. Both +T and +N utilize high-quality gender predictions generated by LLMs. To evaluate the system's robustness, we also examine settings with intentionally incorrect gender information. These are denoted as +WT (Structured Tagging with wrong gender) and +WN (Natural Language with wrong gender).

Metric We adopt the reference-free metric COMET[1] (Rei et al., 2022) as the primary evaluation metric in our experiments to assess quality of translation since no reference of translation is given. Additionally, to evaluate the gender prediction performance of LLMs, we employ accuracy as the metric, treating the task as a binary classification problem.

4.2 Results and Analysis

MT Performance Our main results are presented in Table 1. Overall, the bigger size models offer better results in COMET score, which is consistent with recent works (Xu et al., 2024; Pang et al., 2025), indicate that the reference free COMET metric is suitable to evaluate quality of all systems.

[1]COMET-22 model (*wmt22-cometkiwi-da*)

Moreover, incorporating additional gender information (+N and +T) leads to significant improvements across various LLMs compared to the base systems for all language pairs, with the exception of LLaMA 3.2 3B on En-De and Qwen 2.5 7B on En-Es, where a slight drop in performance is observed. We hypothesize that the relatively small sizes of Qwen 2.5 7B and LLaMA 3.2 3B may limit their ability to effectively interpret prompts, resulting in limited performance gains. Additionally, LLaMA 3.2 3B, having been trained on a more recent dataset, might better capture contextual cues in high-resource languages, i.e. German.

When incorrect gender information (+WT and +WN) is provided, the performance of all models declines significantly across all languages compared to the base models. This indicates that gender information plays a crucial role in helping LLMs interpret inputs and produce accurate translations.

Gender Prediction Accuracy In the first step of our two-step approach, we use LLaMA-3.3-70B-Instruct to predict the gender of each entity in the source sentence based solely on the sentence context. Given the model's strong understanding of English, it achieves a high prediction accuracy of 99.34%, which is consistent with expectations.

5 Ablation study

Since COMET scores show biases in recent reports (Zaranis et al., 2025), we assess whether the observed MT improvements are significant and meaningful in realistic scenarios by employing the LLaMA-3.3-70B-Instruct model as an automatic scorer or judge (Zheng et al., 2023; Li et al., 2024). Comparative results between the base models and those incorporating gender information (+T and +N) are presented in Table 2 and Table 3. An illustrative example is shown in Table 4, with further details provided in Appendix A. Overall, the win rates for systems incorporating gender information (+T and +N) consistently exceed the corresponding loss rates across all languages, with performance gaps ranging from 18% to 40%, demonstrating the effectiveness of incorporating gender information for improving LLM translation quality.

We present an example illustrating the use of gender information in comparison to the base system in Appendix A.

6 Conclusion

This paper explored the use of explicit gender information to reduce gender bias in zero-shot MT and improve the translation performance. We proposed a two-step approach: first, leveraging LLMs to infer the gender of unambiguous entities from context; second, incorporating this information into translation prompts using either **Structured Tagging** or **Natural Language** formats. Comprehensive experiments across five models and four language pairs demonstrate that explicit gender cues consistently improve translation quality, with **Structured Tagging** yielding the most significant gains.

Limitations

In our work, we focus on using explicit gender information to mitigate gender bias at the sentence level in MT, as there is currently no available data to support analysis in broader scenarios such as the document level. However, we plan to develop such resources and conduct further analyses on more realistic and diverse cases in future work.

Bias Statement

Gender in this work refers to binary grammatical gender (masculine and feminine). We define gender bias as the systematic mistranslation of gender-unambiguous entities by LLMs, where incorrect gender assignments occur despite clear contextual cues. Such behavior is harmful because it undermines translation fidelity, introduces stereotypical distortions, and perpetuates inaccurate gender representations in target languages.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

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

Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. 2024. Locating and mitigating gender bias in large language models. In *Advanced Intelligent Computing Technology and Applications*, pages 471–482, Singapore. Springer Nature Singapore.

Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Rada Mihalcea, and Zhijing Jin. 2025. Causally testing gender bias in LLMs: A case study on occupational bias. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4984–5004, Albuquerque, New Mexico. Association for Computational Linguistics.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Anna Currey, Maria Nadejde, Raghavendra Reddy 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 *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY,* *USA, February 7-12, 2020*, pages 7659–7666. AAAI Press.

Joel Escudé Font 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, Florence, Italy. Association for Computational Linguistics.

Anna Farkas and Renáta Németh. 2022. How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points. *Social Sciences Humanities Open*, 5(1):100239.

Harritxu Gete and Thierry Etchegoyhen. 2024. Does context help mitigate gender bias in neural machine translation? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14788–14794, Miami, Florida, USA. Association for Computational Linguistics.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 901–912.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Aida Kostikova, Joke Daems, and Todor Lazarov. 2023. How adaptive is adaptive machine translation, really? a gender-neutral language use case. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 95–97, Tampere, Finland. European Association for Machine Translation.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference, CI 2023, Delft, Netherlands, November 6-9, 2023*, pages 12–24. ACM.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Eva Tsouparopoulou, Dimitris Parsanoglou, Maria Symeonaki, and Giorgos Stamou. 2024. Gost-mt: A knowledge graph for occupation-related gender biases in machine translation. *Preprint*, arXiv:2409.10989.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 300–314, Sheffield, UK. European Association for Machine Translation (EAMT).

Şeyda Portillo-Palma and Sergi Alvarez-Vidal. 2024. Gender bias and contextual sensitivity in machine translation: A focus on turkish subject-dropped sentences. *transLogos Translation Studies Journal*, 7(2):1–28.

Hongye Qiu, Yue Xu, Meikang Qiu, and Wenjie Wang. 2025. Dr. gap: Mitigating bias in large language models using gender-aware prompting with demonstration and reasoning. *arXiv preprint arXiv:2502.11603*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin

Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Shaina Raza, Ananya Raval, and Veronica Chatrath. 2024. MBIAS: Mitigating bias in large language models while retaining context. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 97–111, Bangkok, Thailand. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rachel Rudinger, 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, New Orleans, Louisiana. Association for Computational Linguistics.

Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Gender-specific machine translation with large language models. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 148–158, Miami, Florida, USA. Association for Computational Linguistics.

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

Danielle Saunders 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, Online. Association for Computational Linguistics.

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

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Van-Hien Tran, Raj Dabre, Hour Kaing, Haiyue Song, Hideki Tanaka, and Masao Utiyama. 2025. Exploiting word sense disambiguation in large language models for machine translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 135–144.

Van-Hien Tran, Chenchen Ding, Hideki Tanaka, and Masao Utiyama. 2023. Improving embedding transfer for low-resource machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 123–134.

Van-Hien Tran and Masao Utiyama. 2025. Enhanced zero-shot machine translation via fixed prefix pair bootstrapping. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 10–15.

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

Huy Hien Vu, Hidetaka Kamigaito, and Taro Watanabe. 2024. Context-aware machine translation with source coreference explanation. *Transactions of the Association for Computational Linguistics*, 12:856–874.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and André F. T. Martins. 2025. Watching the watchers: Exposing gender disparities in machine translation quality estimation. *Preprint*, arXiv:2410.10995.

Tao Zhang, Ziqian Zeng, Yuxiang Xiao, Huiping Zhuang, Cen Chen, James Foulds, and Shimei Pan. 2024. Genderalign: An alignment dataset for mitigating gender bias in large language models. *Preprint*, arXiv:2406.13925.

Jieyu Zhao, 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, New Orleans, Louisiana. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Appendix: Case Study

Given the source sentence: "The mechanic fixed the problem for the editor and she charged a thousand dollars.", the pronoun "*she*" should refer to "*The mechanic*". In this context, the gender of "*The mechanic*" should therefore be interpreted as female. Table 4 presents the Italian translation outputs produced by both our two-step approach and a baseline system, using the Qwen2.5-72B-Instruct model.

Among the three candidate translations analyzed, notable differences arise in the accurate representation of gender and the use of appropriate professional terminology. The baseline translation, "Il meccanico ha risolto il problema per la redattrice e lei ha chiesto mille dollari.", fails to reflect the specified female gender of "*The mechanic*", employing the masculine form "Il meccanico". This gender mismatch disrupts linguistic coherence and detracts from the overall fidelity of the translation.

In contrast, our two-step approach explicitly infers the gender of "The mechanic" as female and incorporates this information into the prompting templates (+T and +N). Both variants successfully produce the correct feminine form "La meccanica" in the Italian output.

Further comparison between the two variants reveals subtle distinctions in translation quality. The +N variant, while correctly rendering both professions in the feminine form, opts for "editrice" to translate "editor", a term more closely associated with publishing professionals, potentially introducing an unintended semantic shift. The +T variant, on the other hand, preserves both gender accuracy and role specificity, using "La meccanica" and "la redattrice" to reflect the intended meaning precisely. It also maintains a more natural syntactic flow by avoiding redundant pronoun usage.

Accordingly, the +T variant yields the most accurate and contextually appropriate translation, demonstrating superior handling of both gender agreement and lexical precision in professional contexts.

B Appendix: LLM-as-a-Judge Evaluation

	German		Italian		Portuguese		Spanish	
	W↑	L↓	W↑	L↓	W↑	L↓	W↑	L↓
Llama 3.2 3B	54.86	22.79	40.53	20.2	52.08	15.47	44.89	14.52
Llama 3.1 8B	50.32	20.27	46.78	17.42	46.59	15.03	47.03	13.38
Llama 3.1 70B	49.94	11.55	39.58	18.81	47.54	9.85	42.36	10.16
Qwen 2.5 7B	58.46	20.96	53.16	21.65	57.01	15.21	52.08	15.47
Qwen 2.5 72B	46.21	16.41	52.65	19.00	47.29	14.96	47.41	13.95

Table 2: Win (W) and Lose (L) rates of LLM-as-judge evaluations for systems incorporating gender information (+T) compared to the base models. Results are reported across different language pairs using the LLaMA 3.3 70B Instruct model(Grattafiori et al., 2024).

	German		Italian		Portuguese		Spanish	
	W↑	L↓	W↑	L↓	W↑	L↓	W↑	L↓
Llama 3.2 3B	35.80	21.28	25.57	24.24	41.79	18.06	32.32	15.97
Llama 3.1 8B	54.80	22.16	54.10	19.57	55.30	16.16	54.04	16.41
Llama 3.1 70B	58.08	15.66	53.72	21.09	58.59	10.61	53.66	15.34
Qwen 2.5 7B	60.48	25.06	55.87	27.27	62.06	20.01	51.45	24.94
Qwen 2.5 72B	49.56	24.46	55.37	28.35	49.31	20.09	49.68	21.21

Table 3: Win (W) and Lose (L) rates of LLM-as-judge evaluations for systems incorporating gender information (+N) compared to the base models. Results are reported across different language pairs using the LLaMA 3.3 70B Instruct model

Source Sentence		*The mechanic* fixed the problem for the editor and *she* charged a thousand dollars.
Base		Il meccanico ha risolto il problema per la redattrice e lei ha chiesto mille dollari.
Ours	+T	La meccanica ha risolto il problema per la redattrice e ha chiesto mille dollari.
	+N	La meccanica ha risolto il problema per l'editrice e ha chiesto mille dollari.

Table 4: Translation output examples from English to Italian using the Qwen-2.5-72B-Instruct model, comparing the Baseline and Our Approach.

The prompt template employed for the LLM-as-a-judge evaluation is presented below:

```
You are a professional evaluator for machine translation systems.
Your task is to decide which translation is better overall, based on the following
    criteria:

   1. Faithfulness: Does the translation accurately preserve the meaning of the
       source sentence?
   2. Fluency: Is the translation natural, idiomatic, and well-formed in the target
       language?
   3. Grammar and Style: Is grammar correct and appropriate for the context? Does
       the style match typical usage?
   4. Gender Agreement: Pay close attention to gender agreement between nouns and
       their modifiers (e.g., determiners, adjectives), and between subjects and
       verbs. Ensure that gender and number are used consistently and correctly,
       following the grammar of the target language.
   5. Constraint Compliance: Does the translation respect any additional
       instructions (e.g., required word forms or gendered terminology)?

   Languages:
     - Source language: English
     - Target language: German
   Do not explain your reasoning or clarify. Only choose the better translation
       based on the criteria above .

   Respond with one of the following:
     Answer: Translation A
     Answer: Translation B
     Answer: Both are equally good

  ### Example 1:
  - Source (English): The developer argued with the designer because his idea cannot
       be implemented.
  - Additional information: Use the masculine form for \"designer\"
  - Translation A (German): Der Entwickler stritt mit dem Designer, weil seine Idee
      nicht umgesetzt werden kann.
  - Translation B (German): Der Entwickler stritt mit der Designerin, weil seine
      Idee nicht umgesetzt werden kann.
  - Answer: Translation A

  ### Example 2:
  - Source (English): The developer argued with the designer because she did not
      like the design.
  - Additional information: Use the feminine form for \"developer\"
  - Translation A (German): Der Entwickler stritt mit dem Designer, weil er das
      Design nicht mochte.
  - Translation B (German): Die Entwicklerin stritt mit dem Designer, weil sie das
      Design nicht mochte.\"
  - Answer: Translation B

  ### Example 3:
  - Source (English): The mechanic gave the clerk a present because it was his
      birthday.
  - Additional information: Use the masculine form for \"clerk\"
  - Translation A (German): Der Mechaniker gab dem Angestellten ein Geschenk, weil
      es sein Geburtstag war.
  - Translation B (German): Der Mechaniker gab dem Angestellten ein Geschenk, weil
      es sein Geburtstag war.
  - Answer: Both are equally good

  ### Your turn:
  - Source (English):  {source English sentence}
  - Additional information: {Gender information}
  - Translation A (German):  {Translation of Base system}
  - Translation B (German):  {Translation of system with incorporating gender
      information}
  - Answer:
```