Proceedings of 3rd International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/2024.21411

Gender Prejudice in Large Language Models from the Perspective of Pragmatics-Based on Critical Translation

Yizhuohan Zhang[1,a,†], Minyiyang Zhao[2,b,*,†], Yichun Ma[3,c,†]

[1]School of Humanities, Communication University of China, Beijing, 100024, China
[2]School of Humanities and Social Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
[3]Suzhou North America High School, Suzhou, 215000, China
a. 15839517588@163.com, b. zmyy9699@gmail.com, c. cocoyichunma272@gmail.com
**corresponding author*
[†]These authors contributed equally to this work and should be considered co-first authors.

Abstract: This paper explores the complex interplay between gender bias in Large Language Models (LLMs) and human empirical perspectives, focusing on pragmatic and critical translation approaches. Considering this pressing issue, this work creates a database of 55 Chinese sentences without explicit gender pronounces and tests translation tasks on ChatGPT, Ernie Bot, and Spark Desk platforms. The study reveals that LLMs exhibit varying levels of gender bias, reflecting cultural differences observed in human translators. In the research, the "Gender Stereotype Circle of Large Language Models and Human", a theoretical framework is designed to address these biases through six vital factors and demonstrate the interactions to aid in optimizing how LLMs cope with gender issues, enhancing the output quality, and providing an empirical foundation for the LLMs' training.

Keywords: Large language models, Gender prejudice, Empiricism, Pragmatics, Critical translation

1. Introduction

In recent years, LLMs have created a sensation and made major changes to the world [1]. They greatly grasped the attention of many and spawned research from many perspectives. Popular LLMs like ChatGPT have shown their advanced abilities by helping people deal with problems more productively. These abilities include complex tasks associated with language, such as translation, summarization, conversational interactions, etc. [2]. It can be described as astonishing how diverse their capabilities are.

Nevertheless, new discoveries show that lots of LLMs perpetuate and amplify gender bias and stereotypes when completing certain language tasks. As gender stereotypes have long existed in the use of language [3], this is not something unexpected or strange. Furthermore, this problem is significant because gender bias in LLMs could arise from data collection and algorithm development, reflecting serious problems from the real world [4]. Because of this drawback, LLMs have to be carefully tested to make sure that they give minoritized individuals and communities the same equal treatment [5]. Additionally, ethics regarding social bias has thrown striking issues in natural language processing [6].

Based on the flaw mentioned above, discussions in this area have increased exponentially. For example, researchers explored images generated by DALL-E 2 of gender bias in workplace scenarios [7]. They found that AI images not only replicate the gender stereotypes that exist in workplaces but also reinforce and increase them. Scholars researching on LLM-Generated reference letters also experienced similar results, as they found that LLM-Generated reference letters also depict gender bias and could lead to societal harms like sabotaging application success rates for female applicants [8]. Moreover, stereotypes of gender can occur in AI-generated stories, too. The United Nations Educational, Scientific and Cultural Organization (UNESCO) revealed prevalent gender stereotypes in these stories, as the analyses show that the Large Language Models-in this case, Llama2-significantly highlighted gender asymmetries while creating the stories for boys, girls, women, and men after receiving the prompts given by the authors [9]. All of this research shows that gender prejudice does greatly appear in AI-generated content. They suggest more improvements to be made in the LLMs that exist in our current era. But in the midst of all of these studies, we've noticed a research gap in the gender studies of LLMs-gender stereotypes in translations by LLMs. Namely, the translation from languages that have no distinct gender implications to languages that must operate their sentences with clear gender pronouns.

Given the research gap, the generation qualities of three distinct LLMs will be investigated: ChatGPT-4.0 [10], Ernie Bot-3.5 [11], and Spark Desk-V4.0 [12]. ChatGPT is English-based, leading to better handling of English tasks. Nonetheless, because of a multilingual knowledge base, it can also manage multilingual dialogues with better translation results than machine translation. While Ernie Bot and Spark Desk operate in China and are more adept at handling Chinese tasks with supporting English. Fu and Yang reckon that LLMs system may contain a certain bias on the results of text generation, mainly according to the incompleteness of the dataset, and manual citation may bring in the personal feelings of the citer [13]. Hence, LLMs have different system architectures, providing new ideas for the application of LLM in the study of gender prejudice. Their characteristics are shown in Table 1.

Table 1: Comparison of features of ChatGPT, Spark Desk and Ernie Bot

	Operating Company	**Usable Scope**	**Limitations**	**Cross-language Ability**
ChatGPT	OpenAI	Global	Limited knowledge Unable to ensure authenticity	Equipped
Spark Desk	iFlytek	Mostly in China	Better communication skills in Chinese than in English	Weak
Ernie Bot	Baidu			

In this work, it proposed the research question to discuss the topic of gender prejudice in LLMs from the perspective of pragmatics-based on critical translation. In order to respond to the proposed objectives, this paper aims to shed light on a comparison methodology to unveil differences in gendered languages by humans and three LLMs and proposes a robust framework for elaborating relationships between empiricism and gender bias in humans and LLMs. The insights derived from this study are expected to guide the development of fairer and more equitable language technologies, particularly in gendered languages.

2. Methods

This section presents a research methodology for assessing gender prejudices in LLMs and humans. By comparing translated English and Chinese texts to report the outcomes, the experiment would showcase gender bias in LLMs.

2.1. Design

A dataset involving 55 Chinese sentences were designed for testing gender bias in LLMs and humans. Each study sentence contains one **noun, verb** or **adjective** traditionally perceived as male or female. This is illustrated in the following example:

我的同事很**贤惠**,经常为自己的家人做饭。(My colleague is very **virtuous** and often cooks for oneself family.)

这个人常常**酗酒**来放松自己。(This person often **drinks** heavily to relax oneself.)

那个**士兵**举起了自己的枪。(The **soldier** raised oneself gun.)

The materials are ambiguous with the "自己" (oneself) pronoun which could refer to any gender. Note that language change is dynamic, numerous new things that appear in the development of society are described through language [14]. Therefore, the study has consciously decided to focus on the habit of contemporary speech with gender stereotype words, leading to redundancy in the analysis if all words of whole period are already accounted for.

Not only human but also LLM may pursue different response gender prejudice depending on themselves empiricism. This comparison approach therefore enables a comprehensive analysis of gender representation in human and LLMs, which is crucial for gendered languages.

2.2. Instrument

After designing 55 Chinese sentences and filters, experiments include aligned translations between Chinese and English by humans and LLMs, two comparison groups conducted in July 2024. A visualization of the experiment procedure is presented in Figure 1.

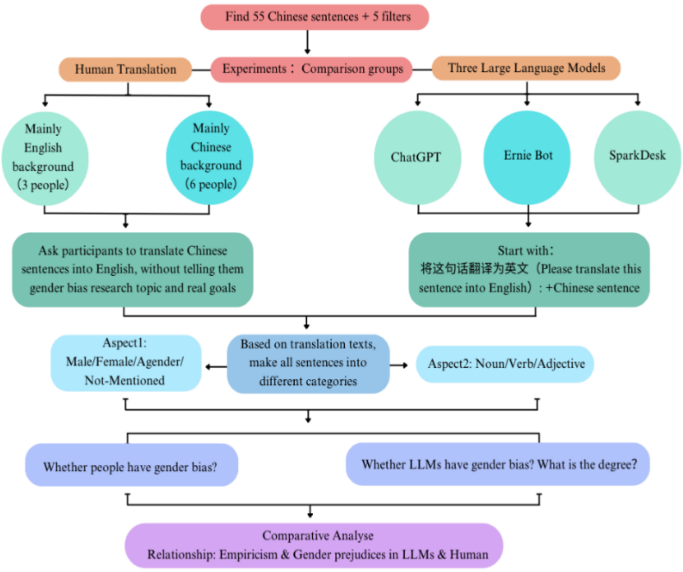
Find 55 Chinese sentences + 5 filters
Human Translation
Experiments：Comparison groups
Three Large Language Models
Mainly English background (3 people)
Mainly Chinese background (6 people)
ChatGPT
Ernie Bot
SparkDesk
Ask participants to translate Chinese sentences into English, without telling them gender bias research topic and real goals
Start with：将这句话翻译为英文（Please translate this sentence into English）：+Chinese sentence
Aspect1: Male/Female/Agender/Not-Mentioned
Based on translation texts, make all sentences into different categories
Aspect2: Noun/Verb/Adjective
Whether people have gender bias?
Whether LLMs have gender bias? What is the degree?
Comparative Analyse
Relationship: Empiricism & Gender prejudices in LLMs & Human

Figure 1: Procedure of experiment

2.2.1. LLMs Translation

The experiment specifically evaluated translation texts on the ChatGPT, Ernie Bot and Spark Desk platforms. The task begins with the identical directive: 将这句话翻译为英文 (Translate this sentence into English). Additionally, it is conducive to diminishing the empiricism by LLMs, through opening a new turn each time translating a new sentence. To evaluate the fairness of LLM output alongside human production, the experiment executed merely one run using the same default settings without any explicit indication. The outcome of 55 sentences. Prior to the analysis, the study carries out coding work to convert the textual data into a format that is suitable for subsequent analysis. During this process, LLMs assign a value of 1 to gender categories, while all other classifications are assigned a value of 0, resulting in an encoded dataset for further examination.

2.2.2. Human Translation

Excluding investigating generation qualities of three LLMs, response human participants include six Chinese native speakers with the second language of English and three vice versa. To avoid cross-trial and cross-task influences, participants were asked to translate sentences from Chinese to English without being told the research theme of gender bias.

2.3. Procedure and Analysis

The translated texts undergo a descriptive investigation of gender-related terms, which are categorized into the following four types: Female (the feminine forms she/her/herself), Male (the masculine forms he/his/him/himself), Agender (the forms they/their/them/themselves and his/her), and None (no pronouns in sentences). Likewise, each phrase is reclassified according to word branches (Nouns, Verbs, and Adjectives). The intention of each procedure is to assess the implication of people and LLMs through conducting multiple comparisons, seeking whether there are significant distinctions in the gender biases conveyed by humans and LLMs.

3. Results

This section introduces the conclusions drawn from our analysis of experimental data. There is a strong possibility that by summarizing the regular experimental results, this work can facilitate the reduction of gender bias in the use of large models in the future.

3.1. Statistical Analysis of LLMs

Studies have shown that different LLMs have different results in determining gender. As shown in Figure 2, when dealing with the same language materials, the most judgment result of ChatGPT is Agender, while that of both Ernie Bot and Spark Desk is Male. Hence, the results show that ChatGPT is more likely to judge gender as Agender, and it has less gender bias in dealing with translation problems. Ernie Bot has almost no Agender and None determination results. In addition, Spark Desk has a lower degree of bias in gender judgment than Ernie Bot.

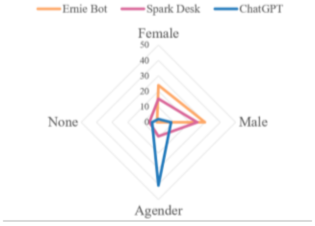

Figure 2: Large language models judgment result radar chart

When faced with keywords of different parts of speech, LLMs also have different judgment tendencies. The study calculates the probability that different large language models judge each part of speech as Male and, similarly, the probability of Female, None, and Agender. Summarizing this rule can facilitate subsequent researchers to improve and reduce gender bias in the translation of large language models. Figure 3 shows that ChatGPT is more inclined to judge nouns as Male and adjectives and verbs as Female. Compared with Spark Desk and Ernie Bot, ChatGPT is not inclined to judge nouns as Female. There seems to be no concept of Agender in Ernie's gender determination program design. All three large language models avoid gender judgment in the process of translation, but ChatGPT and Ernie Bot only appear in the judgment of nouns, and Spark Desk appears in the judgment of nouns, adjectives and verbs. Therefore, the results of this investigation show that Spark Desk is more flexible in avoiding gender judgment in translation.

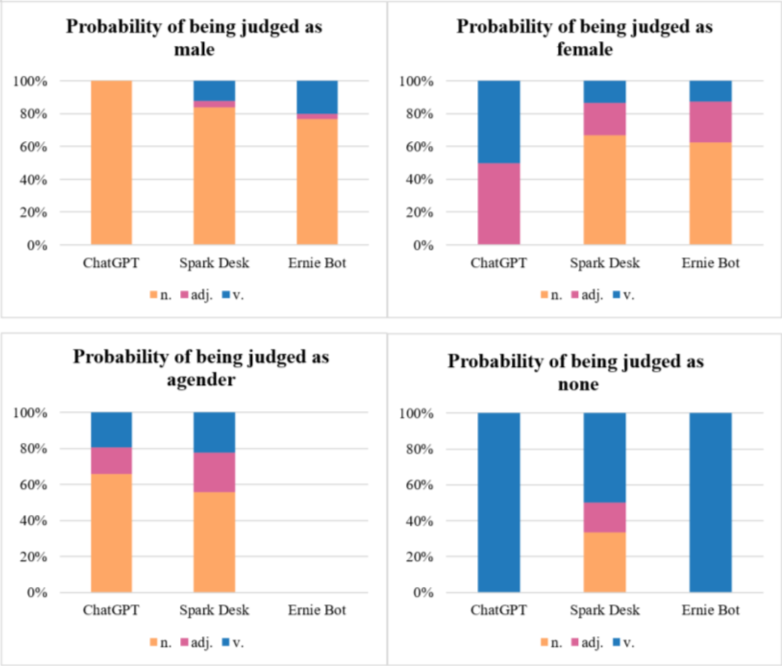

Figure 3: Probability of being judged as Male/Female/Agender/None

3.2. Statistical Analysis of Human

Simultaneously, these sentences are also translated by people from different cultural backgrounds, and the coding data of human translation results are obtained. Studies have shown that individuals from different cultural backgrounds also exhibit different degrees of gender bias in the process of translation. Figure 4 shows that people with English cultural backgrounds are more likely to judge gender as Agender and None in the process of translation. This reflects that people with an English cultural background have lower prejudice than those with a Chinese cultural background. Therefore, what emerges from the results is that different language and cultural environments will affect the degree of gender bias.

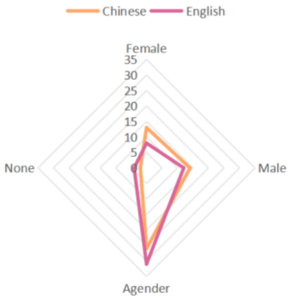

Figure 4: Human judgment result radar chart

3.3. Comparative Analysis of Human and LLMs

Acerbi and Stubbersfield reckon that the LLM reflects a human bias against certain types of content in its production when dealing with tasks [15]. Given the findings, human and LLMs translations have a correlation. Gender bias also exists in people's translation, which confirms that the gender bias of LLMs comes from human beings' bias. The research background shows that ChatGPT is a LLMs developed by people with English cultural backgrounds. The data reveal that ChatGPT displays less gender bias in the process of translation, which is due to the fact that people with English cultural backgrounds have less gender bias.

4. Discussion

4.1. Pragmatics Perspective

From the perspective of pragmatics, this study has many theoretical bases.

Firstly, Chinese and English have different cultural contexts. The study of Hall indicates the high and low context culture theory, Chinese is a high-context language, with less clear coding and more implicit expression [16]. English is a low-context language, so the interpretation of English information is highly dependent on the language itself. For a long time, there have been many gender implications in Chinese words. Therefore, Chinese native speakers capture this point more keenly in the process of Chinese-English translation, forming more gender prejudices. This may be one of the reasons why there is more gender bias in large language models with Chinese background.

The second is the word vector technology in the large language models. Because the word vector technology, there is a rich context behind each individual word, which restricts the understanding of the sentence by the LLMs. It is these contexts that give sentences a pragmatic presupposition and affect the LLMs' judgment of gender. As shown in Figure 5, the word 'pink' is explained by many other words, such as 'bow tie', 'cute', 'rose', 'girl' and so on. These words provide a large language environment for 'pink', making the language model more inclined to judge it as a feminine word.

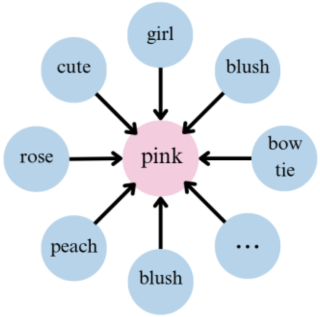

Figure 5: The word vector technology in the large language models

The last is the connection between language and society. Language serves as a reflection of society, mirroring the values and cognitive patterns of a particular culture. Gender differences and gender discrimination are not the natural attributes of language symbols themselves but the reflection of the values and national thinking modes of a specific society in language. Furthermore, language inherently has no natural attribute of gender, but people tend to assign gender to words in the process of long-term use. For instance, the term 'pink' is merely a colour, and it is only when people repeatedly apply it to the description of women in use that the language model makes an error in the gender judgment of 'pink'. The use of LLMs further reinforces the gender bias in the use of the term 'pink'.

4.2. Critical Translation Perspective

Regarding critical translation, Xing and Yang reckon that the study of gender ideology in critical translation mainly explores the interaction in the development of gender concepts in translation [17]. This approach emphasizes the role of translation in shaping and influencing societal gender stereotypes. Moreover, as demonstrated by recent findings, gender differentiation in English language usage is mostly captured through distinct pronouns and other gender-specific tokens [18], which exits a pattern similarly observed in Chinese. Two English-translated outcomes are as follows:

Source Text (ST):我有个老师总称自己为君子。

Target Text (TT) (ChatGPT): I have a teacher who always calls himself a gentleman.

TT (Spark Desk): I have a teacher who always refers to himself as a gentleman.

TT (Ernie Bot): I have a teacher who always calls himself a gentleman.

TT (one of participants): I have a teacher who always refers to themselves as a gentleman.

ST: 那个人的小三最后自己离开了。

TT (ChatGPT): That person's mistress eventually left on her own.

TT (Spark Desk): The person's mistress eventually left by herself.

TT (Ernie Bot): In the end, that man's mistress left on her own.

TT (one of participants): That person's girlfriend finally left.

These gender biases may reflect societal and cultural perspectives in Chinese and English. Chinese terms “君子” and “小三” can both refer to either male or female, but the English counterparts, “gentleman” and “mistress”, carry gender-specific connotations, leading to potential gender stereotypes. This discrepancy highlights a significant challenge in translation, where LLMs may replicate human biases instead of functioning as truly creative and unbiased translators to achieve equivalent expressions. Apparently, LLMs lack the will of a natural human, to a certain extent, they could merely translate source texts from a human perspective without the nuance and flexibility required to fully bridge cultural and gender differences.

Consequently, a new trend in the digital age of LLMs may emerge, namely, embracing critical translation practices to refine and adjust how LLMs handle gender issues in translation. This evolving field would contribute that target texts are not only accurate but also localized, thereby fostering better communication and understanding across LLMs and humans, achieving greater linguistic equivalence and gender sensitivity across languages in the global digital landscape.

4.3. Implication for Practice

Given the findings that gender biases in ChatGPT, Ernie Bot and Spark Desk, are related to human empiricism, it could be proposed a comprehensive model termed the "Gender Stereotype Circle of LLMs and Human". This model aims to cope with gender prejudice in LLMs by incorporating insights from pragmatic and critical translation analyses. It serves as a guiding theoretical framework designed to aid LLMs in better navigating and mitigating issues of gender bias.

The model most noteworthy feature is to showcase interconnections among six components. These components include human empiricism, which reflects the biases and stereotypes that people bring into interactions with the corpus and LLMs, the training data used to develop these models, the algorithms that process and generate responses, and the feedback mechanisms that help refine model outputs. Additionally, the model accounts for the impact of societal norms and the role of ongoing model evaluations in identifying and addressing gender biases.

The exterior structure of the Gender Stereotype Circle even provides a visual representation of how these components interact. This framework not only highlights the complex interplay between human and LLMs elements but also stresses the importance for a multi-faceted approach to reducing gender discrimination in LLMs. By understanding and addressing each component's role in perpetuating biases, developers and researchers can work towards creating more equitable and inclusive language models. Concerning the above mentioned, a visualization of the structure model is presented in Figure 6.

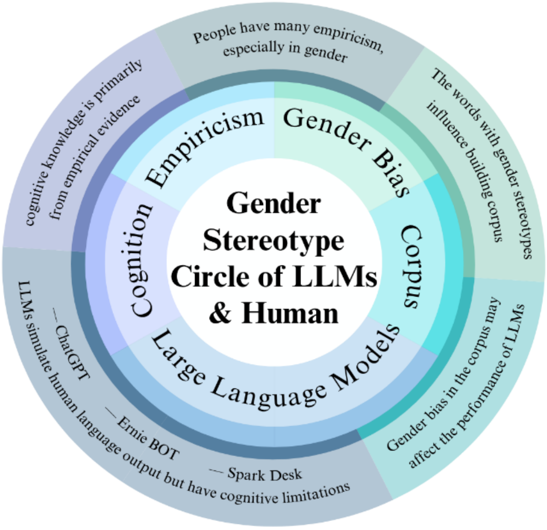
Gender
Stereotype
Circle of LLMs
& Human
Empiricism
Gender Bias
Corpus
Large Language Models
Cognition
People have many empiricism, especially in gender
The words with gender stereotypes influence building corpus
Gender bias in the corpus may affect the performance of LLMs
LLMs simulate human language output but have cognitive limitations
— ChatGPT
— Ernie BOT
— Spark Desk
cognitive knowledge is primarily from empirical evidence

Figure 6: Gender stereotype circle of large language models (LLMs) and human model overall structure

4.4. Limitations

Despite conducting relatively complete research, many limitations still lie beneath. For one, more samples are needed for both the corpus and the translation by humans to further ensure the accuracy and universality of our research. Second, the complete logic behind how LLMs work has details that are still vague. Since it is hard to get access to complete explanations of how they operate, there might be factors that could slightly affect how the models choose gender pronouns during the translation process that we are unaware of. Therefore, for future research, it would be beneficial if we added more samples to refine gender word classifications with his/her aspect and occupation aspects. Additionally, an important question to think about is the significance of using “their” in the translation of LLMs. Although it does represent progress in the reduction of gender bias because “their” is a gender-neutral word, we seldom hear people talk like this in real life. As a result, it is worth thinking if this stands as progress made or a fall behind. Promising results could be discovered in extended studies in the future.

5. Conclusion

This study sheds light on the intricate connection between gender prejudice in Large Language Models and human empiricism, as analyzed through the lenses of pragmatics and critical translation theory. Concerning the analysis of 55 Chinese sentences devoid of explicit gender words, it has been observed that Language Models display varying degrees of gender bias, mirroring the cultural disparities found in human translators. The single most striking observation to emerge from testing three recently released LLMs is that the yielded outcomes are similar among models, implying that these findings may generalize to others available on the market.

Hence, this study proposes the "Gender Stereotype Circle of LLMs and Human" model, a theoretical framework to deal with these issues, addressing six key factors and illustrating each interaction. This model underscores the significance of gender prejudices with cross-languages in the LLM design and training, emphasizing the need for such considerations in future work. Nonetheless, limitations in sample size and understanding of LLM mechanisms necessitate further research to refine gender bias detection and boost LLM fairness, which will ensure the model better reflects natural linguistic patterns while maintaining fairness.

Acknowledgments

Yizhuohan Zhang, Minyiyang Zhao and Yichun Ma contributed equally to this work and should be considered co-first authors.

References

[1] *Euronews. (2023). ChatGPT a year on: 3 ways the AI chatbot has completely changed the world in 12 months. https://www.euronews.com/next/2023/11/30/chatgpt-a-year-on-3-ways-the-ai-chatbot-has-completely-changed-the-world-in-12-months*

[2] *Naveed, H., Khana, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A Comprehensive Overview of Large Language Models. https://doi.org/10.48550/arXiv.2307.06435*

[3] *Malsburg, T., Poppels, T., & Levy, R. (2020). Implicit Gender Bias in Linguistic Descriptions for Expected Events: The Cases of the 2016 United States and 2017 United Kingdom Elections. https://doi.org/10.1177/0956797619890619*

[4] *Manasi, A., Panchanadesaran, S., & Sours, E. (2023). Addressing Gender Bias to Achieve Ethical AI. The Global Observatory. https://theglobalobservatory.org/2023/03/gender-bias-ethical-artificial-intelligence/*

[5] *Kotek, H. Dockum, R. Sun, D. (2023) Gender bias and stereotypes in Large Language Models. In Collective Intelligence Conference (CI '23), November 06--09, 2023, Delft, Netherlands. ACM, New York, NY, USA 13 Pages. https://doi.org/10.1145/3582269.3615599*

[6] *Cho, W., Kim, J., Kim S., & Kim, N. (2019). On Measuring Gender Bias in Translation of Gender-neutral Pronouns. https://doi.org/10.48550/arXiv.1905.11684*

[7] *Garcí-a-UII, F.-J., & Melero-Lázaro, M. (2023). Gender stereotypes in AI-generated images. Profesional De La información, 32(5). https://doi.org/10.3145/epi.2023.sep.05*

[8] *Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K. W., & Peng, N. (2023). "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. https://doi.org/10.48550/arXiv.2310.09219*

[9] *UNESCO. (2024). Challenging systematic prejudices: an investigation into bias against women and girls in large language models. https://discovery.ucl.ac.uk/id/eprint/10188772*

[10] *OpenAI. (2022). Introducing ChatGPT. Retrieved 12 August 2024, from: https://openai.com/blog/chatgpt*

[11] *Baidu. (2023). Ernie Bot. Retrieved 12 August 2024, from: https://yiyan.baidu.com/welcome*

[12] *iFlytek. (2023). Spark Desk. Retrieved 12 August 2024, from: https://xinghuo.xfyun.cn*

[13] *Fu, R., & Yang, X. (2023). Analysis of AIGC Language Models and Application Scenarios in University Libraries. Journal of Library and Information Science in Agriculture. https://link.cnki.net/doi/10.13998/j.cnki.issn1002-1248.23-0406*

[14] *Xu, N., & Yin, Y. (2008). Bias in Linguistic Discrimination—A Study of Anti-Male Language. JOURNAL OF NORTHWEST A&F UNIVERSITY (SOCIAL SCIENCE EDITION), 8(5), 116–119. https://doi-org-s.elink.xjtlu.edu.cn:443/10.3969/j.issn.1009-9107.2008.05.025*

[15] *Acerbi, A., & Stubbersfield, JM. (2023). Large language models show human-like content biases in transmission chain experiments. Proceedings of the National Academy of Sciences of the United States of America, 120:44*

[16] *Hall, E. (1977). Beyond Culture. Anchor Books, New York*

[17] *Xing, J., & Yang, H. (2020). Ideological Explicitation in the Corpus-based Critical Translation Studies: A Review of Introducing Corpus-based Critical Translation Studies. Shandong Foreign Languages Teaching Journal, 41(4), 131–135. https://link.cnki.net/doi/10.16482/j.sdwy37-1026.2020-04-014*

[18] *Derner, E., de la Fuente, S. S., Gutiérrez, Y., Moreda, P., & Oliver, N. (2024). Leveraging Large Language Models to Measure Gender Bias in Gendered Languages.*