Preserving Ambiguity: Prompt Sensitivity in Gender-Neutral Literary Translation by GPT Models

Jin Yim

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

© 2025. Jin Yim. Preserving Ambiguity: Prompt Sensitivity in Gender-Neutral Literary Translation by GPT Models. In Emmanuele Chersoni, Jong-Bok Kim (eds.), *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation* (PACLIC 39), 118-130. Institute for the Study of Language and Information, Kyung Hee University. This work is licensed under the Creative Commons Attribution 4.0 International License.

Preserving Ambiguity: Prompt Sensitivity in Gender-Neutral Literary Translation by GPT Models

Jin Yim
Ewha Womans University
52, Ewhayeodae-gil, Seodaemun-gu,
Seoul 03760, Republic of Korea
jin.yim@ewha.ac.kr

Abstract

This study aims to evaluate how effectively various prompting strategies influence GPT model translations of gender-neutral Korean third-person singular expressions into English, using the queer literary novel *Concerning My Daughter* (Kim, 2017) as a corpus. Specifically, it investigates how the effects of prompting strategies for translating a specific expression have evolved across GPT model versions (GPT-3.5-turbo, GPT-4, and GPT-4o). Through quantitative analyses (BLEU, TER, BERTscore), multivariate statistical techniques (MANOVA, PCA, CA), and qualitative examinations, this research demonstrates significant improvements in translation quality and gender neutrality when context-aware prompts are employed. Meta prompting explicitly emphasizing gender ambiguity was most effective with advanced GPT versions (GPT-4 and GPT-4o), though overly complex prompts did not consistently improve quality metrics. The findings highlight critical limitations in current evaluation frameworks, advocating for specialized criteria and ethical frameworks in AI-generated literary translation.

1 Introduction

As AI-based machine translation (MT) has become commonplace and increasingly applicable to various types of texts, recent studies have begun to explore whether AI can adequately handle sophisticated and subtle literary translation tasks (Alsajri, 2023; Hu & Li, 2023; Li, 2024; Mukti et al., 2024).

This study focuses specifically on the under-researched issue of translating gender-marked references, which poses particular challenges in Korean-to-English literary translation. By examining multiple versions of a single generative AI (GAI) model, the research investigates how sophisticated prompting strategies contribute to producing literary translation outputs increasingly closer to human-level quality.

In particular, gender-neutral language usage is an important concern in MT research, as previous studies have reported gender stereotypes in both neural machine translation (NMT) models (Stanovsky et al., 2019; Vanmassenhove et al., 2018) and large language models (LLMs) (Piazzolla et al., 2024). Research has particularly focused on characterizing and resolving gender bias occurring between specific language pairs such as English and languages with grammatical gender systems, including French and Italian (Ghosh & Caliskan, 2023; Piazzolla et al., 2024; Sant et al., 2024). Although improvements in MT quality have been reported in terms of mitigating gender bias, several challenges still remain unresolved (Piazzolla et al., 2024). Moreover, existing research has predominantly concentrated on binary gender distinctions, while attention toward non-binary gender issues in translation is increasingly gaining prominence (Kostikova, 2023; Yim, 2025; Yu, 2025).

Against the backdrop, this study qualitatively and quantitatively evaluates how the translation outputs of Korean expressions referring to non-binary gender in queer literature vary according to different prompts (zero-shot, context, and meta prompting) and model versions (GPT-3.5 turbo, GPT-4, and GPT-4o) of an LLM, using human translation as the baseline. In doing so, this research aims to contribute to the literature on ethical considerations in AI-generated language usage and literary translation into English. The research questions are specifically formulated as follows:

- How do automated translation quality metric scores for literary sentences containing non-binary gender expressions vary according to prompts and model versions?
- How do gendered or ungendered translation patterns of non-binary gender references differ depending on variations in model versions and prompting strategies?

This study extends previous research (Yim, 2025), which examined prompt sensitivity in a single GPT version (GPT-4o). This comparison allows for an analysis of how prompt sensitivity has evolved across different GPT versions. Furthermore, translating non-binary gender is closely linked to ethical issues in LLM-generated language usage, specifically regarding gender bias, and this comparison will provide insight into how these ethical considerations have evolved across GPT models. Given the observation that non-binary gender has been particularly underexplored in translation and linguistics studies involving Korean (Yu, 2025), this research contributes to expanding existing linguistic scholarship beyond the MT literature by addressing a critical gap in the field.

2 Literature Review

This section first discusses why translating non-binary gender references from Korean to English raises significant ethical and linguistic issues, particularly from a human translation perspective (2.1). It then reviews previous research on how MT has addressed these challenges, highlighting the linguistic implications of these approaches (2.2). Finally, the chapter explores how translation outcomes have been improved across recent GPT models with prompting techniques (2.3).

2.1 Translating Non-binary Gender

The manner in which gender is expressed varies significantly across linguistic systems. English follows a natural gender system, meaning that only nouns and pronouns explicitly indicate a person's gender (Sabato & Perri, 2020). Korean similarly employs the identical gender system; however, its third-person pronoun usage notably differs from English. Korean lacks a fully developed third-person singular pronoun system, instead utilizing demonstratives such as i ("this"), *geu* ("that"), and *jeo* ("that over there") combined with nouns like *saram* ("person"), *i* ("one"), *nom* ("guy"), and *ja* ("individual") to denote third-person referents (Ko & Koo, 2020). Importantly, these combined forms generally do not explicitly indicate grammatical gender, with only a few specialized exceptions, such as *geunyeo* ("she"), which originated from translations of the Japanese equivalent of the English pronoun "she" (Ahn, 2001).

In English, third-person singular pronouns explicitly indicate gender through forms such as "he" and "she," "her" and "his," or "him." However, a significant issue arises because masculine pronouns in English have historically been employed generically to refer to individuals of all genders—a practice increasingly recognized as sexist language (Sabato & Perri, 2020, p. 334). As alternatives, more inclusive forms such as "s/he" and the singular "they" have been adopted. The use of singular "they" has gained formal recognition, appearing in major dictionaries, including the Oxford English Dictionary (2024), where it is explicitly described as suitable for contexts requiring gender neutrality or for referring to individuals who do not identify within the binary gender framework.

In summary, translating third-person singular references from Korean into English necessitates context-based inference to determine whether binary gendered pronouns ("he" or "she") or more inclusive non-binary forms ("s/he" or "they") should be employed. Additionally, it is possible to refer to an individual using a proper noun or a general noun that is either explicitly gendered (e.g., "the man" or "the woman") or non-gendered (e.g., "the person"). This inferential process significantly impacts translation accuracy, intensifying the challenge of maintaining gender-neutral language usage. Especially when gender ambiguity is intentionally encoded in the source text, the translator's choice between gendered or ungendered forms can considerably affect the discourse functions of the translated output, as demonstrated in previous studies such as Aguilar (2023), Ivan (2024), and Yim (2025).

2.2 Issues of Gender Translation in MT

When translating gender-neutral items whose grammatical gender is unknown into gender-marked languages, the process of gender inference becomes essential. If contextual information is insufficient, MT systems either infer and assign a gender or translate in a manner that preserves gender ambiguity. Issues of gender in MT have been primarily explored in relation to gender bias. Previous research has indicated that, when MT systems translate source texts without explicit gender information into target languages that obligatorily mark gender, they tend to exhibit a default bias toward masculine forms (Ghosh & Caliskan, 2023; Piazzolla et al., 2024; Stanovsky et al., 2019; Vanmassenhove et al., 2018). Large language models (LLMs) are reported to be even more susceptible to this bias (Sant et al., 2024). For instance, ChatGPT frequently translates gender-neutral pronouns into explicitly gendered forms such as "he" or "she" (Ghosh & Caliskan, 2023). Similarly, DeepL has shown a notable tendency to overuse the pronoun "he" in backtranslations from Finnish, Estonian, and Indonesian into English. This bias is particularly influenced by sentence context and verbs used, demonstrating high reproducibility across repeated translations (Barclay & Sami, 2024). Such gender bias can significantly impact translation accuracy and discourse effectiveness by introducing gender markers absent from the original text or incorrectly inferring a different gender.

2.3 Improved Translation with Prompting

LLM prompting strategies are effective the translation outputs (Yamada, 2024). Accordingly, various studies have explored prompt design to enhance translation performance across multiple GPT model versions (He, 2024; Peng et al., 2023; Sant et al., 2024; Wang et al., 2023). Several strategies have been found beneficial in enhancing LLM-generated translations: For example, providing contextual information such as detailed translation guidelines and domain-specific knowledge (Peng et al., 2023); assigning a translator persona to the GPT model (He, 2024); and instructing the model to translate at the document level to leverage broader contextual understanding (Wang et al., 2023). It is also known that multi-shot prompts, which provide actual translation examples, improve translation quality more effectively than zero-shot prompts (Sant et al., 2024). Particularly, explicitly instructing the model about the translation's specific purpose and emphasizing the reduction of gender bias has proven effective in decreasing biased outcomes (Sant et al., 2024). Recently, meta-prompting techniques informing the LLM of task details and subsequently allowing it to generate its own prompts have also demonstrated potential for improving translation quality (Suzgun & Kalai, 2024).

However, some studies suggest that simpler prompts might be preferable to overly complex ones. Specifically, concise and effective prompts, such as zero-shot prompts (He, 2024), have been reported to yield better translation quality improvements compared to prompts containing detailed translation briefs and elaborate instructions (Peng et al., 2023). As discussed above, translation outputs significantly vary according to prompt configurations. In terms of GPT model performance, GPT-4 has shown substantial improvements over GPT-3.5 in both translation quality (Jiao et al., 2023; Yan et al., 2024) and other performance metrics (Chen et al., 2024). In particular, GPT-4 reportedly outperforms junior translators but still falls short of the translation quality produced by experienced human translators, showing a tendency toward literal translation (Yan et al., 2024).

In summary, two critical gaps currently exist in GPT models' handling of non-binary gender translations. First, while ChatGPT's translation performance has improved in recent versions, it still has not matched the quality of experienced human translators. Moreover, most prompt-related studies discussed above have primarily focused on single-model versions, leaving uncertainty about how the same prompting strategies affect translation outcomes across different GPT versions. Second, despite extensive prompting research to address gender-related translation problems (Sant et al., 2024), gender biases and translation errors persist. The fact that translation issues remain even in relatively clear-cut cases involving binary gender suggests that the challenges become considerably more complex when translating non-binary gender references. Addressing this significant issue, particularly from the perspective of inclusive language use, calls for future interdisciplinary research evaluating inclusive translation across diverse languages, textual contexts, and GPT model versions. Given these research gaps, the present study aims to examine how translations of sentences containing non-binary gender expressions—requiring heightened gender sensitivity—have evolved across GPT model versions, which currently represent the most widely adopted LLM technology. The findings will indirectly contribute to identifying approaches to accurate gender inference through contextual and grammatical cues, while also addressing methods to preserve the discourse effects of the original text and uphold ethical principles related to gender-neutral language use. Consequently, this study provides significant insights into the current state of AI-generated translation.

3 Methodology

3.1 Corpus

The source text corpus used in this study is drawn from the Korean novel *Concerning My Daughter* (Kim, 2017) by novelist Hyejin Kim. This queer novel portrays the internal struggles experienced by a mother who struggles to accept that her daughter is living with a same-sex partner, narrated from the mother's perspective. Throughout the novel, the mother consistently refers to her daughter's non-binary gender partner using the third-person singular pronoun *geu ae* ("that person/child"), maintaining gender neutrality. From the novel, 183 sentences containing references to *geu ae* were identified and compiled into a corpus. Using this corpus, the study measures how gender neutrality is maintained or altered according to different prompting strategies across multiple GPT model versions. For baseline comparison, this paper used Jamie Chang's English translation (Kim, 2022), which was viewed by critics as "precise and pared-back" renditions of the original narrative "in a careful, balanced way" (West-Knights, 2022). The original text and the corresponding human translation corpus were also utilized in a previous analysis conducted by Yim (2025).

3.2 Process

To investigate translation choices concerning non-binary gender references, each sentence was translated independently using the API. The translation conditions and prompts used were consistent with those employed by Yim (2025) (see Appendix A). The context prompt has the gender and character name, while the simplified form of meta prompt includes the persona, genre, and gender ambiguity instructions.

3.3 Analysis

The analysis was divided into two parts: translation quality (TQ) and gender representation. First, TQ scores were quantitatively evaluated by combining multiple metrics, following the recommendation of Kocmi et al. (2021). Specifically, BLEU (Papineni et al., 2002) was used to assess basic similarity to the human baseline; TER (Snover et al., 2006) was applied to measure the practical effort required for post-editing; and BERTscore (Zhang et al., 2020) was employed to evaluate how effectively the contextual meaning of the original text was captured. To explore how translation quality scores varied according to prompts and GPT model versions, a MANOVA test was conducted. This was followed by PCA (Biber, 1988) to visually illustrate prompt sensitivity showing how prompts make LLM translations deviate from human translations across model versions. Additionally, qualitative analysis involved to identify distinctive patterns and variations.

Second, the analysis examined the representation of non-binary gender in translation outputs (gender representation). The frequency of gendered versus ungendered reference expressions was analyzed across the nine generated corpora and compared against the human translation baseline. Descriptive statistical analysis (chi-square test) and explanatory analysis (Correspondence Analysis; Glynn, 2014) were performed to determine which translations most closely resembled the human baseline in terms of gender-neutrality patterns, while also identifying distinctive translation characteristics according to prompts and GPT model versions.

Python 3.11.8 (July 14, 2025)	**R 4.5.1 (July 20, 2025)**
pandas: 1.5.3	FactoMineR: 2.12
numpy: 1.24.0	factoextra: 1.0.7
scipy: 1.9.3	CA: 0.71.1
matplotlib: 3.6.3	**GPT-3.5 turbo, 4, 4o**
seaborn: 0.11.2	API version: 1.13.3
openpyxl: 3.0.10	Temperature: 0.7
nltk: 3.8.1	Max tokens: 300
sacrebleu: 2.3.1	Top-p: 1.0
bert-score: 0.3.13	Penalty: 0
Okt, konplay 0.6.0	Date: July 3, 2025

Table 1: System and package information

Corpus compilation, data analyses, and statistical testing were conducted using Python and R within the Google Colab environment. Information on the specific models and packages utilized is presented in Table 1.

Corpus (version_prompt)	**ID**	**Token**
Source text	ST	3108
Human translation	HT	2610
GPT-3.5_simple	GPT35_TT1	2948
GPT-3.5_context	GPT35_TT2	2927
GPT-3.5_meta	GPT35_TT3	2868
GPT-4_simple	GPT4_TT1	2794
GPT-4_context	GPT4_TT2	2780
GPT-4_meta	GPT4_TT3	2737
GPT-4o_simple	GPT4o_TT1	2836
GPT-4o_context	GPT4o_TT2	2878
GPT-4o_meta	GPT4o_TT3	2765

Table 2: Corpus size

4 Results

4.1 Translation Quality Metrics

Information about the corpora (number of tokens) generated based on the analysis process presented in Section 3.3 is provided in Table 2.

Table reports corpus-level mean $\pm$ SD ($n = 183$ sentences) for each model-prompt condition. TER is reported as TER inverse so that higher values indicate better quality. Across prompting

TQ per corpus		GPT35	GPT4	GPT4o
BLEU	**TT1**	0.111 ± 0.115	0.123 ± 0.124	0.132 ± 0.127
	TT2	0.119 ± 0.129	0.148 ± 0.147	0.162 ± 0.157
	TT3	0.12 ± 0.136	0.126 ± 0.142	0.152 ± 0.157
	Pairwise	-	-	-
TER inverse	**TT1**	11.228 ± 33.889	12.411 ± 37.514	14.627 ± 34.519
	TT2	16.958 ± 31.414	21.303 ± 32.501	23.459 ± 32.174
	TT3	14.818 ± 32.157	16.318 ± 34.866	23.006 ± 30.915
	Pairwise	-	-	-
BERTScore	**TT1**	0.921 ± 0.024	0.923 ± 0.024	0.925 ± 0.023
	TT2	0.924 ± 0.024	0.929 ± 0.025	0.932 ± 0.025
	TT3	0.924 ± 0.024	0.924 ± 0.026	0.933 ± 0.025
	Pairwise	-	-	tt1 < tt2 *; tt1 < tt3 *

Table 3: TQ description and pairwise comparison

strategies, context prompting (TT2) and meta prompting (TT3) generally produced higher mean scores than simple prompting (TT1) for BLEU, TER_inverse, and BERTScore. TT2-TT3 comparisons were mixed: meta prompting exceeded context prompting only for GPT-3.5 on BLEU and for GPT-4o on BERTScore; in all other cases, TT2 showed the higher mean.

For each model, prompt effects were tested with a Kruskal-Wallis omnibus test followed by Dunn pairwise tests with Holm adjustment. Nonsignificant pairwise contrasts are omitted. Significant differences emerged only for GPT-4o on BERTScore ($tt1 < tt2^*$; $tt1 < tt3^*$; * $p < .05$, Holm-adjusted).

To further assess how prompt types and GPT model versions influenced changes in these metrics, a MANOVA test was performed (Appendix B). Although MANOVA generally requires multivariate normality, the large sample size ($n = 183$) ensures that the analyses conducted remain robust due to the central limit theorem.

The MANOVA results revealed statistically significant effects of prompt type (Wilks' $\lambda = 0.9888$, $F(6, 3284) = 3.08$, $p = .005$) and model version (Wilks' $\lambda = 0.9847$, $F(6, 3284) = 4.24$, $p < .001$) on the three translation quality metrics. In contrast, the interaction between prompt type and model version was not significant (Wilks' $\lambda = 0.9961$, $F(12, 4328.74) = 0.53$, $p = .899$). Although the Wilks' values were close to 1, indicating that the overall effect sizes were small, both prompt type and model version contributed to significant variation in translation quality scores.

Finally, a Principal Component Analysis (PCA) was conducted using the three translation quality metrics scores across corpora to visually represent how prompts make translations deviate from human translations across versions (see Figure 1 and Appendix C for detailed results).

PC1 accounted for comprehensive translation quality, evenly reflecting all three metrics (75.87%), while PC2 captured differences between BERTScore and the other two metrics (12.8%). Compared to BLEU and TER, BERTScore reflects semantic similarity rather than surface-level overlap. The relatively higher BERTScore therefore suggests that, even when lexical realizations diverge, the generated translations retain meaning-level consistency. Together, these two principal components explained 88.67% of the total variance across corpora. Prompt sensitivity was visualized by calculating convex hull areas based on PCA scores. Results indicated that the changes in the areas across prompts remained substantial in GPT-4 (TT1 $\rightarrow$ TT2: -1.575, TT2 $\rightarrow$ TT3: 5.011) and GPT-4o (TT1 $\rightarrow$ TT2: 2.206, TT2 $\rightarrow$ TT3: 2.859), compared to GPT-3.5 (TT1 $\rightarrow$ TT2: 0.893, TT2 $\rightarrow$ TT3: 0.174). This suggests that GPT-4 and GPT-4o exhibited greater sensitivity to prompt changes, reflected in broader variations in translation quality scores compared to the earlier version.

This trend is evident in Example 1 (Appendix D), which shows that score changes increased in conjunction with TT2 and TT3 in GPT-4 and GPT-4o.

4.2 Gender Representation

Corpus	Gendered				Ungendered			
	Female pronoun	Female noun	Male		Proper noun	Pronoun	Noun	
Human	101	21	1	**123**	110	11	2	**123**
GPT35_tt1	78	1	15	**94**	0	26	129	**155**
GPT35_tt2	198	14	1	**213**	25	16	29	**70**
GPT35_tt3	207	2	0	**209**	42	18	17	**77**
GPT4_tt1	51	2	32	**85**	0	19	146	**165**
GPT4_tt2	159	11	2	**172**	112	11	5	**128**
GPT4_tt3	167	5	0	**172**	127	10	5	**142**
GPT4o_tt1	49	10	10	**69**	0	51	134	**185**
GPT4o_tt2	201	12	2	**215**	37	15	5	**57**
GPT4o_tt3	154	4	0	**158**	100	13	3	**116**

Table 4: Frequency of gender representation

Next, to address Research Question 2 regarding gender neutrality, the analysis focused on the frequency of gender reference expressions used across

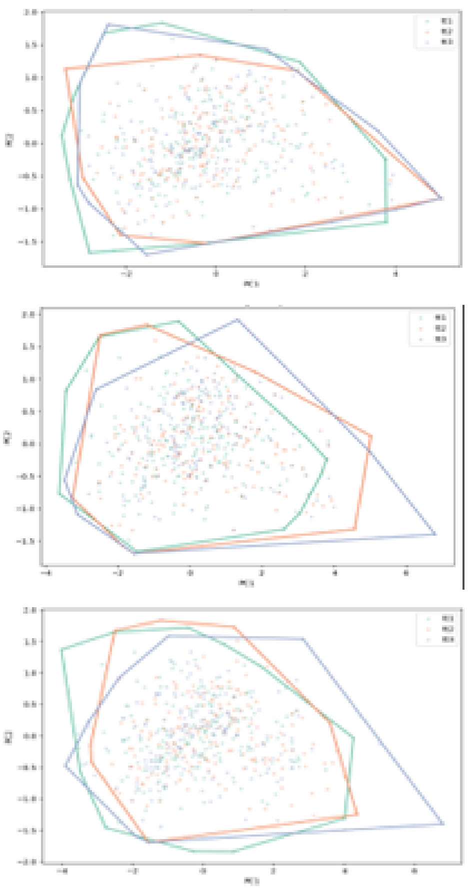
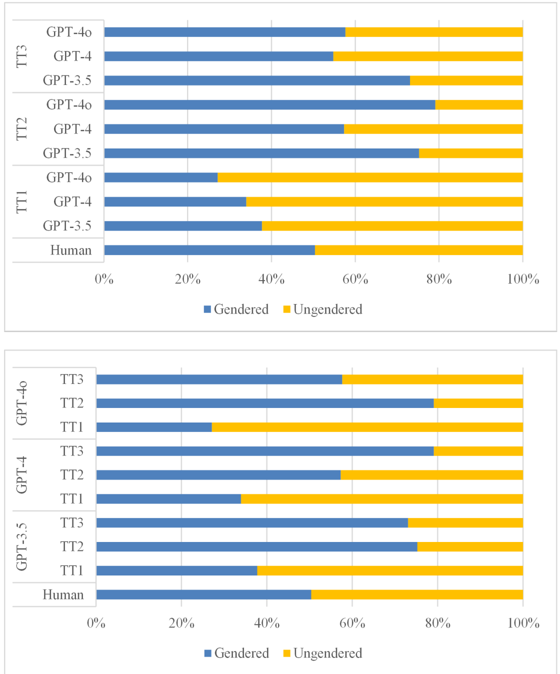


Figure 2: Modelwise prompt effects (GPT-3.5, 4, 4o, top to bottom)



Note: Panel (a) groups the results by prompt type (TT1–TT3), while Panel (b) groups them by model. Both panels display the same dataset in horizontal bar format to improve readability and highlight different comparative perspectives.

Figure 1-a (top), b (bottom). Frequency of gendered vs. ungendered expressions across human and LLM outputs.

each corpus. Following the analytical categories established in previous research (Yim, 2025), frequencies of English translations corresponding to Korean ungendered third-person references were measured. Table 4 presents the corpus-specific frequencies of expressions identified as the most commonly used English translations for the Korean term *geu ae* ("that person/child").

Since some cells contained frequencies of five or less, a chi-square test was conducted using aggregated frequencies of gendered and ungendered expressions across each corpus. For simplicity of comparison, we employed chi-square tests, which capture only overall associations. Such analysis lies beyond the present scope and will be pursued in future work. Although the chi-square test is limited in that it does not capture interaction effects between variables, more advanced approaches such as log-linear modeling would be required for this purpose. The test revealed statistically significant differences among corpora ($\chi^2 = 306.490$, $df = 9$, $p = 1.09E - 60$).

In particular, for the TT1 prompt, which provided no explicit gender-related instructions, the frequency of gendered expressions—indicating binary gender assignment through inference—decreased progressively from GPT-3.5 (94) to GPT-4 (85), and further to GPT-4o (69). This trend suggests increased gender sensitivity in higher GPT model versions (Figure -a).

For context prompting (TT2), which explicitly provided the character name and gender information, the frequency of gendered expressions was highest for GPT-4o (215), followed by GPT-3.5 (213), and GPT-4 (172). Notably, GPT-4 produced the highest frequency of gender-neutral references (128) despite explicit gender information (Figure -a).

In meta-prompting (TT3), which highlighted genre characteristics and emphasized the importance of gender ambiguity, all GPT models showed increased frequencies of ungendered expressions compared to context prompting (TT2). However, the magnitude of this increase varied considerably across models: GPT-3.5 showed an increase of 7 instances (TT2: 70 $\rightarrow$ TT3: 77), GPT-4 an increase of 14 (TT2: 128 → TT3: 142), while GPT-4o exhibited a notably higher increase of 59 (TT2: 57 → TT3: 116) (Table 4, Figure -b).

Notably, GPT-4 and GPT-4o under the TT3 prompting condition exhibited proportions of gendered and ungendered expressions most similar to the human translation baseline. This suggests that prompts incorporating contextual information and explicit instructions emphasizing gender ambiguity were more effective than zero-shot prompting in achieving human-like gender neutrality. However, GPT-3.5 did not achieve human-level performance in representing gender ambiguity.

To further provide explanatory statistical insights, a two-dimensional correspondence analysis (CA) was conducted, offering a detailed visual representation of distances between the human translation and the nine MT-generated corpora concerning specific expressions (Figure 3). Contributions and coordinates for each corpus and expression are provided in Appendix E.

Table 5 shows that the eigenvalue for the first dimension (Dim1) was notably high at 0.554, surpassing the conventional threshold (≥ 0.4) for a robust dimension. Although the eigenvalue for the second dimension (Dim2) was relatively lower at 0.109, it was still meaningful for providing additional explanatory value. Regarding explained inertia, Dim1 accounted for 71.84% of the total variance, and the cumulative inertia up to Dim2 was 86.82%, indicating sufficient reliability and explanatory power for interpreting the analysis results.

Dimension	**Eigenvalue**	**Explained_Inertia**
Dim1	0.554	71.84%
Dim2	0.110	14.28%

Table 5: CA Inertia

Figure 3 shows the correspondence analysis map using the row principal method (FactoMineR + factoextra default). The rows (corpora) are represented in principal coordinates, while the columns (lexical items) are shown as supplementary points. CA results revealed that GPT-4 TT2, GPT-4 TT3, and GPT-4o TT3 were positioned closest to human translation regarding the use of gendered reference expressions. Conversely, as expected, the three TT1 corpora without explicit contextual information diverged significantly from human translation. Despite the provision of context and instructions emphasizing gender ambiguity, GPT-3.5 TT3 still frequently produced gendered translations. This suggests that the effectiveness of prompts emphasizing gender ambiguity increased with the advancement of the GPT model versions, leading to more human-like translations.

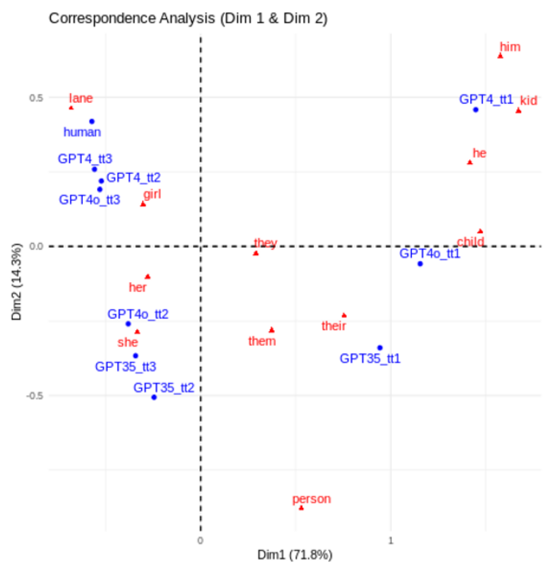


Figure 3: CA Results.

In Dim1, which accounted for most of the corpus differences, the expressions with the highest contributions were *child* (34.8%), *kid* (24.4%), and *Lane* (17.0%). Predictably, the three TT1 corpora—which lacked the explicit character name—were positioned on the opposite end of Dim1. Human translation, GPT-4o TT3, GPT-4 TT2, and GPT-4 TT3 were closely positioned around the proper noun *Lane*, whereas GPT-4o TT2 focused more explicitly on the gender information provided rather than employing the character's name. Additionally, in the TT1 condition, GPT-3.5 favored inclusive forms, whereas GPT-4 defaulted toward masculine forms (Example 2 in Appendix D).

In Example 2, it is also worth mentioning that the pronoun "they" in the human translation could either represent a singular pronoun or, based on context, a plural form referring collectively to Lane and the protagonist's daughter. Given that GPT-3.5 TT1 appeared close to "they" in Figure 3, an examination of the corpus revealed two instances (Examples 3 and 4 in Appendix D) where GPT-3.5 likely employed the gender-inclusive singular form of "they." In contrast, the newer models GPT-4 and GPT-4o translated these references differently, using "the child" (GPT-4 and GPT-4o, Example 3), gendered expressions such as "the kid" and "he" (GPT-4, Example 4), or "the girl" and "she" (GPT-4o, Example 4). These translation choices led general nouns like "child" and "kid" to significantly contribute to distinguishing the corpora along Dim1. Additionally, these tendencies explain why GPT-4o TT1 was closely positioned to explicitly masculine pronouns such as "he" and "him."

Moreover, despite detailed prompting instructions, GPT-3.5 and GPT-4 occasionally failed to correctly infer omitted subjects from the Korean source text or produced misgendered translations (Example 5, Appendix D). In contrast, GPT-4o accurately translated the source meaning across all prompting conditions in the same example, successfully reflecting the intended gender ambiguity through prompting.

5 Discussion and Conclusion

The evaluation of translation quality metrics across three prompting strategies and GPT model versions revealed that quality scores were more strongly influenced by prompts than by model versions. Specifically, context and meta prompting conditions yielded higher scores compared to zero-shot prompting. However, comparisons between context and meta prompting showed mixed outcomes: average scores decreased in six corpora but increased in only three corpora, suggesting that excessively complex prompts may not consistently enhance translation quality. Based on MANOVA results and convex hull visualization from PCA analysis indicated that newer GPT models tended to exhibit greater sensitivity to prompting variations. These findings indicate that prompt-induced improvements in translation quality are negligible for earlier models but emerge with GPT-4o, and primarily on a semantic similarity metric (BERTScore), suggesting that newer model families may better leverage context/persona instructions.

Furthermore, the analysis of frequencies of key gender-marking expressions demonstrated significant differences among corpora in their use of gendered versus ungendered forms. Human translation maintained a balanced 1:1 ratio. All three GPT models under zero-shot prompting tended toward ungendered forms, whereas context prompting resulted in substantially higher proportions of gendered forms. Meta prompting, which provided genre-specific context and explicit instructions emphasizing gender ambiguity with persona, led to increased use of ungendered expressions, with GPT-4o showing the largest increase.

The implications of these findings are as follows. First, although translation quality improved with context prompting compared to zero-shot prompting, adding more detailed information such as genre and gender ambiguity instructions failed to yield further improvements in some corpora, implying two possibilities. On one hand, it partially supports prior research indicating that overly complex prompts may hinder translation quality (He, 2024). On the other hand, it aligns with studies suggesting that providing context can enhance translation performance (He, 2024; Peng et al., 2023; Sant et al., 2024).

Second, the greater effectiveness of meta prompting over context prompting in maintaining gender ambiguity suggests limitations of current automated translation quality metrics in adequately capturing literary translation characteristics. Given the complexity of literary translation evaluation, it is necessary to adopt additional specialized metrics derived from corpus-based human translation studies (Liu et al., 2024) or to develop tailored evaluative frameworks (Zhang et al., 2025). This study particularly emphasizes the importance of incorporating translation criticism approaches—addressing ethical and ideological considerations—into assessments of AI-generated literary translations.

Third, instances where GPT-3.5 translated ambiguous references using singular "they," whereas newer models produced explicitly gendered translations (e.g., "the girl–she," "the child–he"), suggest a concerning trend. Specifically, more recent GPT outputs might not necessarily reflect increased inclusivity regarding non-binary gender. This finding, while preliminary, indicates the need for systematic further analysis. Although this study demonstrates that the latest models can produce gender-neutral translations when explicitly instructed, continued advancements in AI should explicitly aim to promote more inclusive language practices.

Despite the significance of these findings, this study has several limitations. First, the scope was limited to multiple versions of a single GPT model; however, given GPT's widespread use, this limitation is somewhat justified. Future studies should expand the analysis to include diverse models such as DeepL, Gemini and Claude to further examine developments in translating gender-neutral references. Second, the prompts utilized in this study specifically focused on literary translation and gender neutrality. Although narrow, this linguistic focus effectively allows for a detailed exploration of translation guidelines and prompt efficacy. Lastly, due to the creative nature of literary translation tasks, the temperature setting was intentionally increased to 0.7; however, translations were generated through a single iteration. Future research should verify the stability of these findings through repeated translation tasks.

The implications of this study are threefold. First, by highlighting the challenges of generative AI translation between languages with differing gender grammars, this research provides significant insights not only into AI literature but also for translation studies and English writing education. Second, by examining how generative AI has evolved regarding gender grammar, this study underscores the necessity for ongoing research into ethical considerations within AI-generated translations. Lastly, by emphasizing the importance of prompt engineering in contemporary AI models, this study contributes to advancing creative translation research, particularly in literary domains utilizing AI.

Acknowledgments

This study is an extended follow-up to Yim (2025). The source text, baseline human translation corpus, and prompts used in this study are identical to those in the previous study (Yim, 2025) for comparability. The author deeply appreciates the anonymous reviewers' careful reading and insightful comments, which greatly helped improve this article. Any remaining errors or limitations are the author's sole responsibility.

References

Primary Sources

Hye-Jin Kim. 2017. *Concerning My Daughter*. Minumsa, Seoul.

Hye-Jin Kim. 2022. Concerning My Daughter, translated by Jamie Chang. Picador, London.

Secondary Sources

Young-hee Ahn. 2001. Translation of third-person pronouns <He> and <She> into Japanese and Korean: New fictional discourse through translation. *Journal of Japanese Language and Literature*, 17:147–172.

Abdulazeez Alsajri. 2023. Challenges in translating Arabic literary texts using artificial intelligence techniques. *EDRAAK 2023*, (February 2023):5–10. https://doi.org/10.70470/edraak/2023/002

Daniel Herencia Aguilar. 2023. Translating gender ambiguity in literatura: The case of *Written on the Body*. Skokpos, 12:137–160.

Peter J. Barclay and Ashkan Sami. 2024. Investigating markers and drivers of gender bias in machine translations. In Proceedings of the 2024 IEEE International Conference on Software Analysis, Evolution and Reengineering *(SANER)*, March 12, 2024. IEEE, Rovaniemi, Finland, pages 455–464. https://doi.org/10.1109/saner60148.2024.00054

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How is ChatGPT's behavior changing over time? *Harvard Data Science Review*, 6(2). https://doi.org/10.1162/99608f92.5317da47

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*, August 08, 2023. ACM, Montréal QC Canada, pages 901–912. https://doi.org/10.1145/3600211.3604672

Dylan Glynn. 2014. Correspondence analysis: Exploring data and identifying patterns. In *Human Cognitive Processing*, Dylan Glynn and Justyna A. Robinson (eds.). John Benjamins Publishing Company, Amsterdam, pages 443–485. https://doi.org/10.1075/hcp.43.17gly

Sui He. 2024. Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 316–326, Sheffield, UK. European Association for Machine Translation (EAMT).

Kaibao Hu and Xiaoqian Li. 2023. The creativity and limitations of AI neural machine translation: A corpus-based study of DeepL's English-to-Chinese translation of Shakespeare's plays. *Babel*, 69(4). https://doi.org/10.1075/babel.00331.hu

Alexandra Maria Ivan. 2024. Gender identity in translation: The impossibility of transposing non-binary characters into Romanian. *Forum for Contemporary Issues in Language and Literature*, 4:27–43. https://doi.org/10.34739/fci.2023.04.03

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. Retrieved from http://arxiv.org/abs/2301.08745

Young-Kun Ko and Bon-Kwan Koo. 2018. *Urimal Munbeopron [Korean Grammar Theory]*. Jipmoondang, Paju, Korea.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 478–494.

Aida Kostikova. 2023. Gender-neutral language use in the context of gender bias in machine translation (a review literature). *JCAL-JCAL*, 1(July 2023):94–109. https://doi.org/10.33919/JCAL.23.1.5

Qi Li. 2024. Bridging languages: The potential and limitations of AI in literary translation—A case study of the English translation of *A Pair of Peacocks Southeast Fly*. *AHR*, 8(1):1–7. https://doi.org/10.54254/2753-7080/8/2024091

Cuilin Liu, Se-Eun Jhang, Homin Park, and Hyunjong Hahm. 2024. A corpus-based multilingual comparison of AI-based machine translations. *kjell*, 24(January 2024):257–276. https://doi.org/10.15738/kjell.24..202404.257

Muhammad Abdee Praja Mukti, Muhamad Trian Maulana, Kharisma Nur Rohmah, Fourus Huznatul Abqoriyyah, and Andang Saehu. 2024. Effectiveness of artificial intelligence usage as translation medium among English literature student of UIN Sunan Gunung Djati Bandung. *JEEF*, 4(2):107–112. https://doi.org/10.29303/jeef.v4i2.683

Oxford English Dictionary. 2024. *Oxford English Dictionary Online*. Oxford University Press. Retrieved July 11, 2025, from https://www.oed.com/

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4390455

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia*, Pennsylvania, USA, pages 311–318. https://doi.org/10.1002/andp.19223712302

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2024. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *Hermes – Journal of Language and Communication in Business*, 63:209–225. https://doi.org/10.7146/hjlcb.vi63.137553

Bruna Di Sabato and Antonio Perri. 2020. Grammatical gender and translation: A cross-linguistic overview. In *The Routledge Handbook of Translation, Feminism and Gender,* pages 363–373.

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. https://doi.org/10.48550/ARXIV.2407.18786

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers,* pages 223–231.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. https://doi.org/10.48550/arXiv.1906.00591

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding. https://doi.org/10.48550/arXiv.2401.12954

Eva Vanmassenhove, Christian Hardmeier, and

Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 3003–3008. https://doi.org/10.18653/v1/D18-1334

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. https://doi.org/10.48550/arXiv.2304.02210

Imogen West-Knights. 2022. In a Korean Best Seller, Women Have Biases, but No Options. The New York Times. https://www.nytimes.com/2022/09/26/books/review/concerning-my-daughter-kim-hye-jin.html

Masaru Yamada. 2024. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. https://doi.org/10.48550/arXiv.2308.01391

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. GPT-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. https://doi.org/10.48550/arXiv.2407.03658

Jin Yim. 2025. Gender ambiguity in human and AI translations: "That Person" in *Concerning My Daughter*. Korean Journal of English Language and Linguistics, 25:2–34. doi:10.15738/kjell.25..202507.912

Huijae Yu. 2025. A linguistic approach to queer translation – Focusing on non-binary identity. *The Journal of Translation Studies*, 26(2):79–103. https://doi.org/10.15749/JTS.2025.26.2.003

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. https://doi.org/10.48550/arXiv.2410.18697

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. https://doi.org/10.48550/arXiv.1904.09675

A Prompts (Yim, 2025)

Corpus	Prompt
TT1(Zero-shot)	"Each translation request must be treated independently, without remembering or referring to previous requests. Do not retain memory or context. Please translate each Korean sentence into English."
TT2 (Context)	"Each translation request must be treated independently, without remembering or referring to previous requests. Do not retain memory or context. In this context, '그 애' refers to Lane, a woman. Please translate each Korean sentence into English."
TT3 (Meta)	"Each translation request must be treated independently, without remembering or referring to previous requests. Do not retain memory or context. You are a professional literary translator with deep sensitivity to gender identity, emotional nuance, and queer relationships. Please translate each Korean sentence into fluent, expressive, and natural English. Preserve ambiguity where appropriate, and maintain the rhythm, tone, and intimacy of the original text. In this context, '그 애' refers to Lane, a woman. Please translate each Korean sentence into English."

B MANOVA Results

Effect	**Wilks' λ**	**F**	**df1**	**df2**	**p**
Prompt	0.9888	3.0814	6	3284	0.0052
Model	0.9847	4.2355	6	3284	0.0003
Prompt × Model	0.9961	0.5265	12	4328.74	0.8991

C PCA Results

Explained variance and loadings

Principal Component	**Metrics**	**Value**
PC1	Explained Variance (%)	75.87
	BLEU_z Loading	0.579
	TER inverse z Loading	0.582
	BERTScore_z Loading	0.571
PC2	Explained Variance (%)	12.8
	BLEU_z Loading	-0.495
	TER inverse z Loading	-0.305
	BERTScore_z Loading	0.813

Centroid coordinates by Corpus

Corpus	**PC1**	**PC2**
GPT3.5 TT1	-0.311	-0.036
GPT3.5 TT2	-0.107	-0.019
GPT3.5 TT3	-0.142	-0.009
GPT4 TT1	-0.189	-0.019
GPT4 TT2	0.197	-0.005
GPT4 TT3	-0.088	-0.037
GPT4o TT1	-0.064	-0.001
GPT4o TT2	0.369	0.033
GPT4o TT3	0.336	0.095

Convex hull area changes across prompting

Version	TT1	Δ	TT2	Δ	TT3
GPT-3.5	18.161	0.893	19.054	0.174	19.228
GPT-4	19.932	1.575	18.356	5.011	23.368
GPT-4o	17.718	2.206	19.924	2.859	22.784

D Examples from Corpora

Example 1: ID #731
ST: 그 애가 잠시 고개를 들어 나와 눈을 맞춘다.
HT: She raises her head for a moment, meets my gaze.

GPT-3.5

TT1: That child briefly raises their head and meets my eyes.

TT2: She raises her head for a moment and meets my eyes (BLEU: 0, TER: -20, BERTScore: +0.0062).

TT3: She raises her head for a moment, meets my gaze (BLEU: 0, TER: 0, BERTScore: +0.005).

GPT-4

TT1: The kid lifts his head for a moment and makes eye contact with me.

TT2: Lane momentarily raises her head and makes eye contact with me (BLEU: 0.0081, TER: +80, BERTScore: +0.0086).

TT3: Lane lifts her head for a moment, locking eyes with me (BLEU: 0.0081, TER: +80, BERTScore: +0.0093).

GPT-4o

TT1: The child briefly lifts their head and makes eye contact with me.

TT2: Her eyes, which had been lingering at a point past me, meet mine (BLEU: -0.02 TER: +40, BERTScore: +0.0093).

TT3: She briefly lifts her head and meets my eyes (BLEU: -0.02, TER: +40, BERTScore: +0.0093).

Example 2: ID #773
ST: 그 애가 오고 난 후 내가 가장 처음 건넨 말은 그것이다.
HT: Those were the first words I said to her after they moved in.
GPT-3.5 TT1: The first thing I said after he/she came was that.
GPT-4 TT1: The first thing I said to him after he came was that.
GPT-4o TT1: The first thing I said after that child arrived was that.

Example 3: ID #669
ST: 그 애는 차분하게 자신의 이름을 말하고 찾아온 용건을 설명한다.
HT: Lane gets up in surprise and evenly states name and purpose of visit.
GPT-3.5 TT1: They calmly state their name and explain the reason for their visit.
GPT-4 TT1: The child calmly states his name and explains the purpose of his visit.
GPT-4o TT1: The child calmly states their name and explains the reason for their visit.

Example 4: ID #3693
ST: 주방에서 과일을 씻던 그 애는 젠이 잠든 줄로만 알았다고 했다.
HT: Lane, who was washing fruit in the kitchen, thought she was asleep.
GPT-3.5 TT1: The person washing fruits in the kitchen said that they thought Zen had fallen asleep.
GPT-4 TT1: The kid, who was washing fruits in the kitchen, said he thought Zen was asleep.
GPT-4o TT1: The kid, who was washing fruits in the kitchen, said he thought Zen was asleep.
GPT-3.5 TT2: She said that she thought Jen had fallen asleep just because she was washing fruit in the kitchen.
GPT-4 TT2: The girl, who was washing fruit in the kitchen, said she thought Zen had fallen asleep.
GPT-4o TT2: The girl, who was washing fruit in the kitchen, said she thought Jen was asleep.
GPT-3.5 TT3: The one washing fruit in the kitchen said she thought Jen had fallen asleep.
GPT-4 TT3: Washing fruit in the kitchen, Lane said she had thought Jen was asleep.
GPT-4o TT3: Lane, who was washing fruit in the kitchen, said she thought Jen was asleep.

Example 5: ID # 1350
ST: 오셨어요. 그 애다.
HT: Hi, you're home, says Lane.
GPT-3.5
TT1: You came. That kid.
TT2: She has arrived. (Ø)
TT3: She's here. It's her.
GPT-4
TT1: He's here. That's him.
TT2: She's here. It's Lane.
TT3: She's here. It's Lane.
GPT-4oS
TT1: You're here. It's that kid.

TT2: You're here. It's her.
TT3: You're here. It's Lane.

E Detailed CA Results

Corpus	Coordinates		Contributions	
	Dim 1	Dim 2	Dim 1	Dim 2
human	-0.572	0.419	5.351	14.450
GPT35_tt1	0.943	-0.340	14.672	9.591
GPT4_tt1	1.447	0.459	34.681	17.513
GPT4o_tt1	1.155	-0.058	22.348	0.286
GPT35_tt2	-0.245	-0.506	1.121	24.082
GPT4_tt2	-0.522	0.219	5.407	4.804
GPT4o_tt2	-0.381	-0.260	2.604	6.092
GPT35_tt3	-0.342	-0.367	2.219	12.827
GPT4_tt3	-0.558	0.259	6.489	7.018
GPT4o_tt3	-0.530	0.191	5.108	3.337

Coordinates & Contributions (Corpus)

Gender	Word	Coordinates		Contributions	
		Dim 1	Dim 2	Dim 1	Dim 2
Gendered	she	-0.332	-0.287	4.612	17.291
	her	-0.278	-0.101	3.811	2.552
	him	1.577	0.638	4.150	3.417
	he	1.417	0.280	5.093	1.001
	girl	-0.303	0.141	0.429	0.470
	woman	-0.680	0.463	17.086	39.860
Ungendered	lane	0.374	-0.281	0.280	0.798
	they	1.471	0.050	34.846	0.199
	their	1.672	0.455	24.469	9.094
	them	0.530	-0.878	1.728	23.842
	child	-0.332	-0.287	4.612	17.291
	kid	-0.278	-0.101	3.811	2.552
	person	1.577	0.638	4.150	3.417

Coordinates & Contributions (Gender Expression)