



Gender reference in bidirectional Korean-English LLM translation: Same-sex partner reference in queer literary texts^{*}

Jin Yim

(Ewha Womans University)

Yim, Jin. 2026. Gender reference in bidirectional Korean-English LLM translation: Same-sex partner reference in queer literary texts. *Linguistic Research* 43(Special Edition): 41-76. Explicitation has often been treated as a translation universal, but its language-dependence remains contested. Focusing on overt gender marking, this study examines whether gender information added in one translation direction is omitted in the reverse direction. A bidirectional corpus was compiled from comparable queer literary texts in which references to a female protagonist's same-sex partner are gender-neutral in Korean but gender-marked in English. OpenAI GPT-5.4, Google Gemini 3.1 Flash-Lite, and Anthropic Claude Sonnet 4.6 translated the source texts under four prompt conditions across three runs, with published human translations as a baseline. Outputs were manually coded and analyzed using binary mixed-effects logistic regression. Korean-to-English translations tended to add gender marking, whereas English-to-Korean translations tended to omit it, with both patterns aligning with target-language conventions. Prompt effects varied by model, and protagonist information exerted a stronger influence than explicit gender-retention instructions. Human translations showed the same directional tendency but accommodated target-language conventions more moderately. These findings suggest that gender explicitation and implicitation are shaped by language-specific conventions and that LLM evaluation for literary translation should consider discourse-sensitive choices beyond grammatical and propositional accuracy. (Ewha Womans University)

Keywords gender explicitation/implicitation, large language models, prompting, queer literature, Korean-English translation

^{*} I am deeply grateful to the anonymous reviewers for their valuable insights, which greatly helped improve this study.

1. Introduction

Large language models (LLMs) have been deployed across multiple domains of human activity, with applications that are potentially reshaping everyday practices of text generation (Liang et al. 2025). Although translation was not the original purpose of LLMs, growing evidence suggests that they can outperform commercial systems based on neural machine translation (NMT) by allowing users to adjust outputs through prompting (e.g., Hendy et al. 2023; He 2024; Savoldi et al. 2024). Their use has also expanded from simple informational texts to increasingly sophisticated genres (e.g., Gao et al. 2024; Guan 2024; Karaban and Karaban 2024; Zhang et al. 2024). With the popularization of LLMs, numerous models have been released, and a wide array of studies now investigates which of these, and which prompting strategies, best serve particular tasks from the user's perspective. Since commercial LLMs are continuously updated and widely adopted in authentic user contexts, examining how prompt effects shift across them is crucial for both understanding real-world language use and assessing the stability of user strategies.

Against this backdrop, the present article consolidates and systematizes the analytical framework of Yim (2025a, 2025b, 2025c) and extends the unidirectional corpus into bidirectional translations between Korean and English to examine how the source text's gender marking manifests in translations generated by three proprietary commercial LLMs under four prompting conditions. The investigation focuses specifically on the overt addition or suppression of gender information as one manifestation of explicitation and implicitation (Klaudy and Károly 2005) in the translation of queer literature, where gender manifestation functions as a key discourse device central to literary sensibility (von Flotow 1997; Rose 2018).

2. Background

As Yim (2025a, 2025b, 2025c) has noted, there is a pressing need to focus on less-explored linguistic features in LLM-generated translations of literary texts. Building on these previous studies, the present study investigates personal references more systematically across an extended corpus, prompts, and models in order to capture broader patterns relevant to the gender-sensitive use of language.

2.1 Personal references in translation

A common typology contrasts grammatical gender languages with pronominal gender languages: in the former, such as French, nouns and their agreement targets are obligatorily gender-marked, whereas in pronominal gender languages such as English, gender appears mainly in personal pronouns (Corbett 1991). Research on gender bias has so far concentrated largely on grammatical gender systems, addressing cases in which mismatches in agreement or lexical choice create visible problems in translation (von Flotow 1997; Zhou et al. 2019; Dawkins et al. 2024; Lardelli 2024; Savoldi et al. 2024; Vanmassenhove 2024). Pronominal gender systems, by contrast, have received comparatively little attention, presumably because they appear less gendered at the grammatical level. Yet even in pronominal gender languages, certain referential expressions may carry significant discourse effects, and this is particularly the case for the English-Korean pair, whose two languages have markedly different pronoun systems (Yim 2025a, 2025b, 2025c).

By the Middle English period, English had shed most of its nominal and adnominal inflection and become a "pronominal gender system" (Curzan 2003: 20; cf. Corbett 1991), in which gender surfaces only in the third-person singular pronouns; for human referents, the *he/she* opposition makes gender obligatory. Such a referent therefore carries different gender information depending on the form used to realize it. Personal pronouns such as *she/he/her/him/his* and gender-marked nouns such as *the girl* or *the woman* specify gender, whereas *that person*, general nouns such as *child* or *kid*, singular *they/their/them* (Krauthamer 2021), and proper names leave it unexpressed. Korean, by contrast, is a pro-drop language in which subjects and pronouns are frequently omitted (Han 2006). Because of its less developed pronoun system, a demonstrative-plus-noun expression such as *geu ae* ('that child') or *jeo saram* ('that person') surfaces overtly to represent third-person singular reference (Goh and Koo 2018).

This typological difference has been reported to cause difficulties in both Korean–English translation (Kim et al. 2019) and L2 English pronoun use (Song 2022). When the gender-ambiguous Korean third-person singular expression *geu ae* is translated into English, it may be rendered in an ungendered form, such as singular *they* or a general noun like *child*, *kid*, or *person*; or in a gendered form, such as *she/her*,

the girl, or the woman (Yim 2025a).

Conversely, when she/her is translated into Korean, gender may be left unmarked because Korean commonly drops subjects and pronouns, or retained through gender-marked expressions such as *geu nyeo* or *geu yeoja*. The status of *geu nyeo* has itself been debated. S. W. Lee (2012) traces the pronoun to translation from English and argues that it should therefore be avoided. S.-B. Lee (2013) accepts this account of its origin but holds that the form may nonetheless gain acceptance as the language changes, noting that younger speakers are increasingly familiar with it. Nevertheless, because of both this controversy over translationese and the structural differences between the two languages, she/her is often omitted in English-to-Korean translation (S. W. Lee 2012: 206–207). Thus, depending on translation direction, differences in gender marking and Korean's pro-drop tendency can lead either to the loss of gender information or to its addition in the target text.

Such shifts are related to the general tendency for translated texts to become more explicit than their source texts (Baker 1996). Not all explicitation is of the same kind, however: Klaudy (2009) distinguishes obligatory explicitation, which stems from grammatical constraints, from optional explicitation, which arises from discourse or stylistic choices. Klaudy and Károly (2005: 16), for example, classify Hungarian *belépett* translated into English she entered as obligatory grammatical explicitation through gender specification, and the reverse as implicitation through gender neutralization. They note, however, that explicitation in one direction is not always counterbalanced by implicitation in the other, even in the addition and omission of grammatical items (2005: 17). Becher (2010, 2011), by contrast, rejects explicitation as a translation universal altogether, attributing such directional differences instead to syntactic and communicative contrasts specific to the language pair. The two accounts differ in what they attribute explicitation to, but not in whether it occurs. What follows from either is that its scope and direction are an empirical matter, to be established for a particular language pair rather than assumed. Throughout this article, the terms gender explicitation/implicitation are used in a restricted operational sense to refer, respectively, to source-relative increases and decreases in overt gender marking. This operationalization does not encompass the broader concept of explicitation or the discourse-level recoverability of gender information.

Whether gender marking is added or omitted, these divergent renderings are all acceptable in terms of propositional accuracy. Given cross-linguistic differences in

gender marking (Corbett 1991), such shifts, from a feminist translation perspective, amount to more than lexical choices and may carry ideological and discursive implications (von Flotow 1997; Irshad 2022; Lardelli 2024). In literary contexts, choices to overtly mark or suppress gender through pronouns and other gendered expressions can themselves function as translational interventions (Miletich 2012), and the presence or absence of overt gender marking can shape discourse interpretation (Rossi 2021). Examining how AI-generated translations handle this dimension reveals whether LLMs preserve or alter gender marking, offering insight both into the sociolinguistic impact of such outputs and into their potential utility for translating creative texts. Building on this discussion, recent studies have examined how LLMs realize gender in their output.

2.2 Gender and personal reference in LLM translation

Among the models studied to date, the GPT series has received the most attention, with versions 3.5 (Karaban and Karaban 2024) and 4 (Jiao et al. 2023) reportedly producing output comparable to that of human translators.

Most evaluations, however, have relied on automated metrics such as BLEU or COMET, with a primary emphasis on overall quality (e.g., Jiao et al. 2023). While such measures capture accuracy, they reveal little about how specific linguistic features shape discourse. Recent work has begun to address this gap, both through case studies of LLM translations of literary works (Karaban and Karaban 2024; Zhang et al. 2025) and through the inclusion of resources such as the Genderqueer Test Suite in large-scale MT evaluations like WMT24 (Friðriksdóttir 2024; Kocmi et al. 2024).

Personal reference is one such feature. Its translation, in which gender-neutral expressions in the source text may be rendered into gender-specific forms in the target text, lies at the intersection of research on gender bias in NMT systems (Cho et al. 2021; Savoldi et al. 2021; Piazzolla et al. 2023; Vanmassenhove 2024), LLM translation outputs (Dawkins et al. 2024; Savoldi et al. 2024), and LLM-generated language more broadly (Hossain et al. 2023; Bas 2024; Tang et al. 2025). Gender bias in LLMs has been shown to vary with model and version (Friðriksdóttir 2024; Tang et al. 2025) as well as with prompting conditions (He 2024; Savoldi et al. 2024).

Unlike NMT systems, LLMs are distinguished by their responsiveness to prompts,

which exert a decisive influence on output (Hendy et al. 2023; He 2024; Savoldi et al. 2024). Prompts commonly used in translation include zero-shot with context, task-specific instructions, and persona (He 2024; Ngo et al. 2025), few-shot (Savoldi et al. 2024), and Chain-of-Thought (Sant et al. 2024). Prior studies suggest that contextual or instructional cues mitigate gender-related issues more effectively (Savoldi et al. 2024; Tang et al. 2025). Other findings, however, indicate that zero-shot prompts still yield higher overall translation quality than prompts involving translation briefs or author personas, even as assigning a translator persona has been found to enhance performance (He 2024).

Prompt effects also differ across models and versions. Under baseline zero-shot prompting, gender neutrality was low in GPT-3.5 (Vanmassenhove 2024) and GPT-4 (Savoldi et al. 2024; Ngo et al. 2025), and gender-neutral pronouns were used inaccurately (Hossain et al. 2023). Substantial improvements were nonetheless reported in GPT-4 under few-shot prompting (Savoldi et al. 2024) and in GPT-4o under zero-shot prompting (Ngo et al. 2025). Yim (2025c) similarly found that gender ambiguity was better handled in GPT-4o than in earlier GPT versions, though Yim (2025a) showed that even GPT-4o fell short of human-level realization of gender ambiguity despite the use of a gender-ambiguity prompt. Claude and Gemini were also competitive, with Claude-3.5 Sonnet achieving the highest accuracy in one comparison (Ngo et al. 2025) and Claude-3.5 rated most accurate in translating gendered they, followed closely by GPT-4 (Friðriksdóttir 2024). Recent work on gender identity inference reports that GPT-4o (Ngo et al. 2025; Tang et al. 2025) and Claude-4-Sonnet (Tang et al. 2025) handle gender-neutral pronouns better even in zero-shot settings, with further gains under few-shot conditions (Tang et al. 2025). Yet gains from prompting (He 2024; Savoldi et al. 2024; Ngo et al. 2025) or from newer models (Friðriksdóttir 2024; Tang et al. 2025) do not necessarily yield sociolinguistically acceptable translations (Savoldi et al. 2024), and challenges persist with non-binary gender and neopronouns (Tang et al. 2025).

Prior work is limited in three respects. Most investigations have centered on high-resource languages (e.g., Bas 2024) or English-Chinese (e.g., Jiao et al. 2023), leaving Korean underexplored in terms of gender rendering. Findings are also difficult to compare, as each study targets different models, prompts, and language pairs; analysis under unified conditions is therefore valuable for establishing comparability and enabling replication as models are updated. Finally, recent research has largely

relied on synthetic datasets to evaluate gender bias (e.g., Bas 2024; Tang et al. 2025), which may not capture naturally occurring referential patterns and discourse functions found in published literary texts. The present study therefore draws its test items from authentic literary works while controlling their presentation at the sentence level for experimental purposes.

Against this backdrop, the present study asks whether gender explicitation is language-dependent: if gender is explicitated in one translation direction, is it left implicit in the other? To address this, I compiled a bidirectional corpus of comparable queer literary texts centering on a female protagonist's same-sex partner. The Korean-to-English component draws on source texts previously shown to elicit gender explicitation (Yim 2025a, 2025b, 2025c); the English-to-Korean component runs in the opposite direction, with gender marked explicitly in the source. Published human translations serve as a baseline. Each source text was translated by three commercially available LLMs (GPT, Gemini, and Claude) under four prompt conditions, guided by the following sub-questions:

1. To what extent do Korean-to-English LLM translations explicitate gender information that is neutral in the Korean source texts?
2. To what extent do English-to-Korean LLM translations retain or implicate overt feminine gender marking that is explicit in the English source texts?
3. Do patterns of gender explicitation and implicitation vary across models?
4. To what extent does the provision of protagonist information alter the treatment of gender, and do explicit instructions — to preserve source-text marking or to produce a natural literary translation — modify that effect?
5. Taken together, do model and prompt effects differ by translation direction, indicating that gender explicitation and implicitation are mediated by the interaction of language direction, model behavior, and prompting?

3. Methodology

This study designed a bidirectional corpus consisting of a KE (Korean-to-English) subcorpus and an EK (English-to-Korean) subcorpus. The corpus comprises sentences containing third-person singular references in Korean and English queer literature,

together with their translations generated by three LLMs under four prompting conditions, and published human translations as a baseline. The following sections describe the corpus design, translation generation, and analysis procedures.

3.1 Corpus

The bidirectional corpus includes two queer source texts: a Korean novel, *Concerning My Daughter* (Kim 2017), and an English novel, *Oranges Are Not the Only Fruit* (Winterson 1997) first published in 1985. Both works depict a daughter raised in a conservative family and the conflict that arises when she enters a same-sex relationship, with the daughter's relationship with a female partner becoming a site of familial opposition that reveals tensions surrounding gender, sexuality, and social norms. The analysis focuses on personal references denoting the female protagonist's same-sex partner: the KE corpus comprises 169 Korean source sentences containing the third-person singular reference *geu ae*, while the EK corpus contains 102 English sentences using third-person singular *she/her*. Both also include a baseline human translation, into English (Kim 2017/2022) and into Korean (Winterson 1985/2024) respectively, alongside the LLM-generated translations. The KE source text and its human translation were used in Yim (2025a, 2025b, 2025c), which examined a single model's (GPT) previous version.

3.2 Translation generation

Translation data were generated on June 29 and 30, 2026, using three commercially available LLMs accessed via API calls: OpenAI GPT-5.4, Google Gemini 3.1 Flash-Lite, and Anthropic Claude Sonnet 4.6. Using the API ensured that each sentence was translated independently, without influence from previous prompts or user history. Sentence-level translation was adopted as a controlled experimental design to isolate the effects of prompt-supplied information. Under P1, the referent's gender was intentionally left unspecified to examine how models translate gender-neutral reference when gender is unknown. From P2 onward, the referent's gender was explicitly provided, allowing the study to examine translation choices made when gender was known rather than inferred from surrounding context.

To maximize comparability, generation settings were standardized across the three models, with temperature set to 0.2 and `max_output_tokens` to 1024; parameters unavailable or not user-configurable for a particular model were left at their default values. API calls were executed in Python 3.13.5. Statistical analysis was conducted in R 4.6.0 using the *lme4* and *emmeans* packages on July 3 to 5, 2026.

To examine the influence of prompting on the realization of gender marking, four prompt conditions were designed for each translation direction, none of which included exemplars. The first (baseline) requested a straightforward translation, which supplied no gender information and thus served as a baseline for comparison; the second (name) additionally provided the protagonist's name and gender; the third (retention) instructed the model to preserve the degree of gender marking found in the source text; and the fourth (literary) requested a natural literary translation while also providing the protagonist's name. The complete prompt texts are given in Appendix 1. To examine output variability, each source sentence was translated three times under each condition, yielding 36 translations per source sentence (3 models $\times$ 4 prompts $\times$ 3 runs). Appendix 2 summarizes basic information on these translations and on the baseline human translation.

Sentences containing third-person references were manually coded when the referent was clearly identifiable as the daughter's same-sex partner and the source and target sentences were sufficiently well aligned to permit sentence-level comparison. Sentences involving translation failure were excluded. The final dataset comprised 169 Korean-to-English (KE) sentences and 102 English-to-Korean (EK) sentences.

For the KE corpus, each third-person reference to Lane, the daughter's female same-sex partner, was first coded binarily for gender explicitness. Gender-marked forms received a code of 1: the feminine pronouns *she* and *her*; the masculine *he*, *him*, and *himself*, produced through misgendering; and gendered nominals such as *that girl* or *that woman*. Gender-neutral renderings received a code of 0, including the name *Lane*, which does not itself signal gender, general nouns such as *that person*, *kid*, *child*, and *partner*, and singular *they/them*. For the EK corpus, a code of 0 was assigned where the source text's gender marking was retained, through expressions such as *geu nyeo*, *geu yeoja*, or *geu sonyeo*. A code of -1 was assigned when the overt gender marker disappeared, whether through replacement with a gender-neutral expression such as *geu ae*, substitution with the female character's name *Melanie*, or complete omission.

An explicitness score (ES) was computed as the number of overtly gender-marked references in the target sentence minus the corresponding number in the source sentence. ES therefore operationalizes the degree of source-relative change in overt gender marking between the source and target sentence. Positive and negative values are referred to as gender explicitation and gender implicitation, respectively, in this restricted operational sense. The measure does not assess the overall semantic explicitness of gender information or its recoverability from the wider discourse. For the regression analysis, ES was dichotomized as zero versus non-zero, so that the outcome variable indexes whether any source-relative shift in overt gender marking occurred, rather than how many individual markers were affected.

To assess coding reliability, approximately 5% of source sentences in each corpus was sampled, and all 37 translations of the sampled sentence (222 cases from the EK corpus and 296 from the KE corpus) were independently coded by a second coder, a professional Korean-English translator, using the same guidelines. Agreement was complete on the direction of the shift; disagreement arose only in the counting of references, affecting 31 cases. These concerned primarily whether possessive forms such as *her* should be counted separately and whether pronouns referring to individuals other than the target referent should be included. Following discussion, the guidelines were refined so that all relevant possessive occurrences were counted, whereas only renderings referring to the target individual entered the analysis. The 31 cases were then recoded accordingly.

3.3 Analysis

The analysis was carried out in two phases. To assess inter-run stability, the three outputs generated for each sentence under the same model and prompt condition were compared at the level of ES direction. For Korean-to-English translation, ES values greater than zero were coded as gender explicitation, whereas ES values of zero were coded as no shift. For English-to-Korean translation, ES values below zero were coded as gender implicitation, whereas ES values of zero were coded as no shift. Sentence-level three-run response patterns and exact agreement rates were calculated for each model–prompt condition. Chance-corrected agreement was assessed using Gwet’s AC1 (Gwet 2008) and Fleiss’ kappa (Fleiss 1971). Because the binary categories could be unevenly distributed, Gwet’s AC1 was treated as the primary chance-corrected

agreement measure, with Fleiss' kappa reported as a complementary statistic.

Mixed-effects logistic regression was then used to model the probability of explicitation or implicitation (Hosmer et al. 2013). Separate models were fitted for each translation direction using the lme4 package in R (Bates et al. 2015). Model, prompt, their interaction, and run were included as fixed effects, with GPT and P1 as reference levels; run was treated as a fixed effect because three levels are too few to support stable variance estimation. A random intercept for source sentence accounted for repeated observations and sentence-level heterogeneity. The model structure was specified as $shift \sim model * prompt + run + (1 | sent_id)$. Human translations were excluded from the regression, since they are not associated with a model or prompt condition, and are reported separately as a baseline.

4. Results

The analysis is presented in three parts, with ES serving as the primary measure. The first examines the overall distribution of gender shifts, including a comparison with the human baseline, and the stability of translation choices across the three runs. The second focuses on the KE corpus, comparing ES patterns across models and prompt conditions. The third presents the corresponding analysis of the EK corpus.

4.1 Overview of source-relative explicitness score

Inter-run stability was assessed first (Table 1). Agreement in the occurrence of gender explicitation or implicitation ranged from moderate to very high across model-prompt conditions, with Fleiss' kappa between 0.453 and 0.937 and Gwet's AC1 between 0.498 and 0.968. AC1 exceeded kappa in several conditions, likely because it is less sensitive to imbalanced distributions between sentences involving a gender shift ($ES \neq 0$) and those involving none ($ES = 0$). Exact agreement rates and the relative frequencies of the two outcome categories were therefore considered alongside both coefficients.

All three models were highly reproducible under P1, with the exception of GPT in the KE corpus. Reproducibility tended to decline as prompts became more elaborate, and GPT was generally less stable than Gemini and Claude under P2-P4.

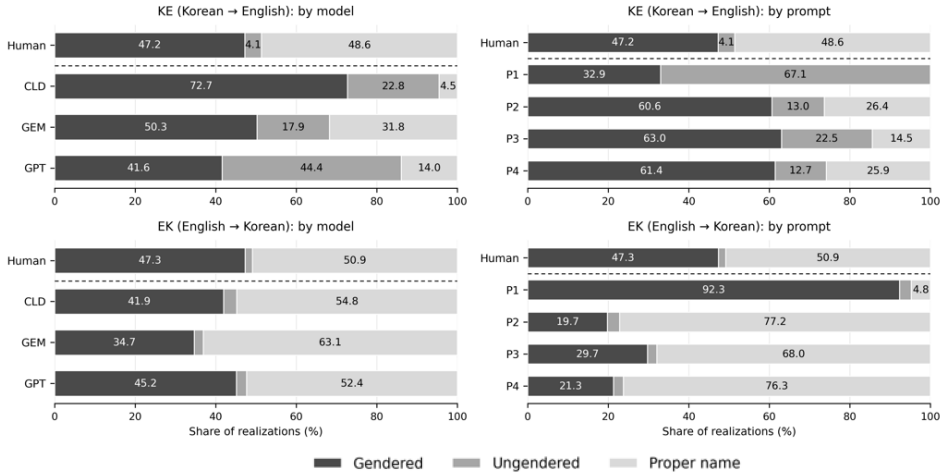
Table 1. Inter-run stability of ES

Model	Prompt	% of ES = 0	% of ES $\neq$ 0	% of Exact Agreement	Fleiss kappa	Gwets AC1
Korean to English Translation						
CLD	P1	67.06%	32.94%	95.86%	0.9375	0.9505
	P2	15.98%	84.02%	96.45%	0.9118	0.9676
	P3	10.26%	89.74%	92.90%	0.7429	0.942
	P4	15.78%	84.22%	94.67%	0.8664	0.9516
GEM	P1	50.30%	49.70%	92.31%	0.8974	0.8974
	P2	52.66%	47.34%	92.90%	0.9051	0.9056
	P3	31.36%	68.64%	91.72%	0.8717	0.903
	P4	48.13%	51.87%	91.72%	0.8894	0.8897
GPT	P1	87.18%	12.82%	81.66%	0.4529	0.8425
	P2	40.43%	59.57%	60.95%	0.4595	0.4977
	P3	60.95%	39.05%	66.27%	0.5277	0.5709
	P4	46.94%	53.06%	62.72%	0.5011	0.5048
English to Korean Translation						
CLD	P1	78.43%	21.57%	100.00%	1.000	1.000
	P2	3.59%	96.41%	98.04%	0.811	0.986
	P3	24.51%	75.49%	91.18%	0.841	0.907
	P4	5.23%	94.77%	97.06%	0.802	0.978
GEM	P1	75.49%	24.51%	100.00%	1.000	1.000
	P2	2.61%	97.39%	98.04%	0.743	0.986
	P3	4.25%	95.75%	97.06%	0.759	0.979
	P4	1.96%	98.04%	100.00%	1.000	1.000
GPT	P1	78.43%	21.57%	96.08%	0.923	0.960
	P2	10.46%	89.54%	87.25%	0.546	0.895
	P3	19.93%	80.07%	77.45%	0.529	0.779
	P4	17.32%	82.68%	81.37%	0.566	0.826

Inter-run stability was also generally higher in the EK corpus than in the KE corpus. This may suggest that outputs are more reproducible when gender is overtly marked in the source, whereas an underspecified source, as in the KE corpus, leaves the model to choose whether to preserve gender neutrality or introduce overt gender marking, thereby increasing variation across runs. The difference may also reflect the near-ceiling rates of implicitation in the EK corpus, which leave little room for variation between runs.

We first examine the overall translation patterns in the KE and EK corpora, each comprising 37 sets of translations (Appendices 3 and 4, respectively). Figure 1 presents the distribution of realization types for the expressions under analysis (Korean *geu*

ae and English *she/her*) across models and prompting conditions, distinguishing gendered forms (personal pronouns and gender-marked nouns) from gender-neutral forms and proper names.



Gendered = gender pronouns + gender-marked nouns, Ungendered = neutral pronouns + neutral nouns, Values below 4% are unlabeled.

Figure 1. % of translation choices by corpus, model, and prompt

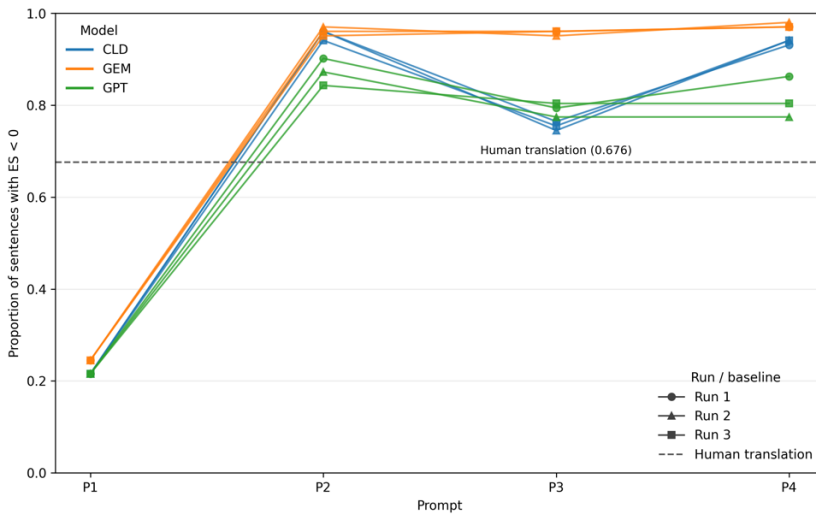
In the KE corpus, gendered forms generally occurred more frequently in the LLM-generated translations than in the human baseline (47.2%), with the exception of GPT (41.6%) and P1 (32.9%). In the EK corpus, by contrast, they fell below the human baseline (47.3%) across all models and prompting conditions except P1 (92.3%). LLM translations thus tend to add gender information in the Korean-to-English direction while omitting it in the English-to-Korean direction.

The three-way breakdown further reveals different translation strategies. In the KE corpus, the human translator and GPT produced similar proportions of gendered forms (47.2% and 41.6%), yet resolved the remaining cases in opposite ways: the human translation relied almost entirely on the proper name (48.6%, against 4.1% gender-neutral forms), whereas GPT favored gender-neutral nouns and pronouns (44.4%, against 14.0% proper names). Claude and Gemini fell between these poles, with Gemini closest to the human pattern in its use of the proper name (31.8%). In the EK corpus, no such divergence appeared: the human translation and all three

models converged on the proper name (50.9% and 52.4-63.1%), with gender-neutral forms accounting for no more than 3.3% in any condition.

For the ES analysis, all translations were coded according to the procedure described in Section 3.3, yielding 6,253 Korean-to-English and 3,774 English-to-Korean translations, including the human baseline. ES was then aggregated by model, prompt, and run, as shown in Figures 2 and 3 for the KE and EK corpora respectively.

In the Korean-to-English data, gender explication occurred in 46.2% of the sentences in the human baseline translation, whereas the three models displayed distinct response patterns across prompts. Overall, explication was comparatively low for GPT and Claude under P1 and tended to increase once protagonist information was supplied, whereas Gemini remained comparatively high across conditions. The extent of these model- and prompt-related differences is tested statistically in Section 4.2.

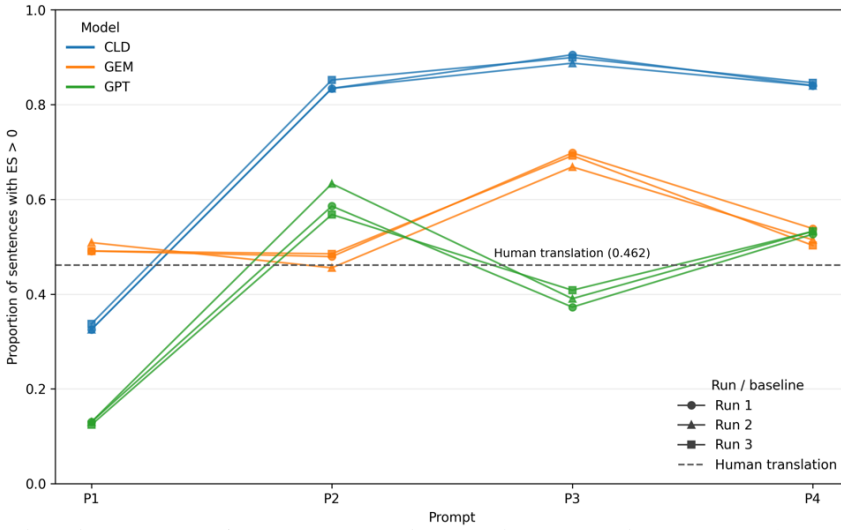


Points show the proportion of source sentences whose translation received a positive ES ($ES > 0$). The dashed line marks the human baseline (46.2% of sentences).

Figure 2. % of positive ES in KE corpora

In the opposite direction (Figure 3), English-to-Korean translations showed a different pattern. Since this direction was designed to examine gender implicitation, the analysis focused on renderings of she/her referring to Melanie. All three models

showed comparatively low implicitation under P1, followed by substantially higher rates once protagonist information was supplied. Descriptively, the retention condition (P3) reduced implicitation relative to P2 for GPT and Claude but not for Gemini, while implicitation remained high under the literary condition (P4). These model- and prompt-related differences are tested statistically in Section 4.3.



Points show the proportion of source sentences whose translation received a negative ES ($ES < 0$). The dashed line marks the human baseline (67.6% of sentences).

Figure 3. % of negative ES in EK corpora

4.2 Prompt and model effects in KE corpus

To test the model- and prompt-related differences summarized descriptively in Section 4.1, the binary mixed-effects logistic regression models described in Section 3.3 were applied separately to the Korean-to-English and English-to-Korean data.

First, the analysis of the KE corpus revealed a significant model $\times$ prompt interaction ($\chi^2(6) = 416.30$, $p < .001$), indicating that the effect of prompt on gender explicitation differed by model. Main effects are therefore reported only as conditional effects at the reference level, and all comparisons below are simple effects within each level of the moderating factor. The model was fitted to 6,084 observations from 169

source sentences; the random intercept for source sentence had a variance of 1.59 ($SD = 1.26$), indicating substantial variation across sentences in the baseline probability of explicitation. The run effect was non-significant ($\chi^2(2) = 0.02$, $p = .993$). Predicted probabilities from the regression are given in Table 2, estimated for each model-prompt combination and averaged over the three runs, while Tukey-adjusted post hoc tests in Appendices 6 and 7 compare models within each prompt and prompts within each model.

Table 2. Predicted probabilities of gender shift

Model	Prompt	KE Probability (explicitation)	95% CI	EK Probability (implicitation)	95% CI
GPT	P1	.080	.057–.110	.053	.022–.122
GPT	P2	.632	.565–.694	.981	.956–.992
GPT	P3	.365	.302–.432	.953	.901–.979
GPT	P4	.549	.480–.616	.962	.918–.983
GEM	P1	.504	.436–.573	.094	.042–.199
GEM	P2	.473	.405–.542	.997	.992–.999
GEM	P3	.740	.682–.791	.997	.991–.999
GEM	P4	.533	.464–.601	.999	.995–1.000
CLD	P1	.288	.233–.350	.053	.022–.122
CLD	P2	.892	.857–.920	.996	.990–.999
CLD	P3	.937	.912–.955	.937	.868–.971
CLD	P4	.894	.859–.921	.994	.984–.998

* Predicted probabilities are population-level estimates evaluated at the random-intercept mean; they may diverge from observed rates where between-sentence variance is large.

Under the baseline prompt (P1), where the isolated Korean source segments did not specify the referent's gender, all three pairwise model contrasts reached significance. Gemini and Claude both exceeded GPT ($OR = 11.77$, $z = 13.43$, $p < .001$; $OR = 4.67$, $z = 8.33$, $p < .001$), and Claude fell below Gemini ($OR = 0.40$, $z = -6.20$, $p < .001$). Predicted probabilities ranged from .080 for GPT to .504 for Gemini, indicating that, when gender was intentionally underspecified, the three models differed substantially in their default tendency to introduce overt gender marking.

Once the proper name was supplied under the name prompt (P2), GPT and Claude increased sharply over their P1 rates ($OR = 19.85$, $z = 16.11$, $p < .001$; $OR = 20.51$, $z = 17.35$, $p < .001$), whereas Gemini did not differ from its own P1 rate ($OR = 0.88$, $z = -0.87$, $p = .821$). Thus, the P2 effect differed by model: Gemini fell below GPT ($OR = 0.52$, $z = -4.47$, $p < .001$), while Claude reached the highest probability

of the three (.892; vs. GPT, OR = 4.82, $z = 9.39$, $p < .001$).

The retention instruction was assessed by comparing P3 with P2 because both conditions supplied the proper name. Since the Korean source expressions are gender-neutral, compliance with the instruction entails a decrease in explicitation relative to P2. GPT's decrease was significant (OR = 0.33, $z = -7.44$, $p < .001$), confirming its comparative responsiveness to the instruction, though its predicted probability under P3 (.365) remained well above its P1 level (.080). For Gemini and Claude the corresponding contrasts were significant in the opposite direction (OR = 3.17, $z = 7.76$, $p < .001$; OR = 1.79, $z = 2.87$, $p = .022$), indicating that their elevated rates under P3 were not a carry-over from P2 but a further increase attributable to the prompt itself; Claude's predicted probability reached .937, the highest value in the KE design. Between-model differences were widest under this condition, with Claude exceeding GPT (OR = 25.85, $z = 17.03$, $p < .001$). Table 3 illustrates the divergence: under P3, GPT returned the referent to the child, reproducing its own P1 output, while Gemini and Claude retained she.

Table 3. Gender explicitation examples in KE corpus

ST K008 run 1	여느 때처럼 그 애는 수궁하듯 고개를 끄덕이지만 그대로 돌아서진 않는다. (As usual, <u>that child</u> nods as if in agreement, but does not turn away just like that)			
HT	Lane nods as always, but does not walk away.			
Model	P1	P2	P3	P4
GPT	(ES: 0) As usual, <u>the child</u> nods as if in agreement, but <u>does not</u> simply turn away.	(ES: 1) As usual, <u>Lane</u> nods as if in agreement, but <u>she doesn't</u> turn away right then.	(ES: 0) As usual, <u>the child</u> nods as if in agreement, but <u>does not</u> turn away just like that.	(ES: 2) As always, <u>she</u> nods as if in agreement, but <u>she does not</u> turn away just yet.
Gemini	(ES: 2) As usual, <u>she</u> nods as if in agreement, but <u>she does not</u> turn away.	(ES: 1) As usual, <u>Lane</u> nods as if in agreement, but <u>she does not</u> turn away.	(ES: 1) As usual, <u>Lane</u> nods as if in agreement, but <u>she does not</u> turn away.	(ES: 1) As usual, <u>Lane</u> nods as if in agreement, but <u>she does not</u> turn away.
Claude	(ES: 0) As usual, <u>the kid</u> nods as if in agreement but <u>doesn't</u> turn away just like that.	(ES: 1) As usual, <u>she</u> nods as if in agreement but <u>doesn't</u> turn away just like that.	(ES: 1) As usual, <u>she</u> nods as if in agreement but <u>doesn't</u> turn away just like that.	(ES: 2) As <u>she</u> always did, <u>she</u> nodded as if in agreement, but <u>did not</u> turn away.

Whether a directive toward literary naturalness altered how gender was handled is a separate question. Relative to P3, all three models differed significantly, but in opposite directions: GPT increased (OR = 2.12, $z = 5.14$, $p < .001$), while Gemini and Claude decreased (OR = 0.40, $z = -6.18$, $p < .001$; OR = 0.57, $z = -2.78$, $p = .028$), reversing for each model the shift that the retention instruction had produced. Between models, GPT and Gemini converged under this prompt (OR = 0.94, $z = -0.43$, $p = .901$), the only KE condition in which they were statistically indistinguishable, while Claude remained above both (vs. GPT, OR = 6.93, $z = 11.53$, $p < .001$; vs. Gemini, OR = 7.38, $z = 11.89$, $p < .001$). Against P2, however, none of the three differed significantly (GPT, OR = 0.71, $z = -2.39$, $p = .078$; Gemini, OR = 1.27, $z = 1.67$, $p = .343$; Claude, OR = 1.02, $z = 0.09$, $p = 1$), as illustrated in Table 3. Thus, the literary-naturalness instruction produced no statistically reliable change in gender handling beyond that associated with the provision of the proper name alone.

4.3 Prompt and model effects in EK corpus

The same analytic procedure was applied to the English-to-Korean data to test the model- and prompt-related differences summarized descriptively in Section 4.1. The model $\times$ prompt interaction was significant ($\chi^2(6) = 64.31$, $p < .001$), whereas the effect of run was not ($\chi^2(2) = 1.96$, $p = .376$), as in the KE corpus. The model was fitted to 3,672 observations from 102 source sentences; the random intercept for source sentence had a variance of 12.39 (SD = 3.52), considerably larger than in the KE model, indicating that implicature varied much more across sentences in this direction.

Under the baseline prompt (P1), no pairwise model contrast reached significance (Gemini vs. GPT, OR = 1.85, $z = 1.64$, $p = .228$; Claude vs. GPT, OR = 1.00, $z = 0.00$, $p = 1$; Claude vs. Gemini, OR = 0.54, $z = -1.64$, $p = .228$), with predicted probabilities clustered between .053 and .094. This was the only condition in either translation direction where the three models could not be distinguished.

Supplying the proper name produced significant increases for all three models (all $p < .001$), with predicted probabilities rising to .981 for GPT, .997 for Gemini, and .996 for Claude, as shown in Table 2. Because implicature was rare under P1,

the baseline odds were extremely low and the odds ratios for these contrasts are correspondingly inflated; they are interpreted here as evidence of a substantial difference between P1 and P2 rather than as absolute effect sizes. At P2, Gemini and Claude both exceeded GPT (OR = 7.02, $z = 4.42$, $p < .001$; OR = 5.38, $z = 4.11$, $p < .001$), while not differing from each other (OR = 0.77, $z = -0.51$, $p = .864$).

Table 4. Gender impicitation examples in EK corpus

ST E034 run 1	We were quiet, and I traced the outline of <u>her</u> marvellous bones and the triangle of muscle in <u>her</u> stomach.			
HT	(ES: -1) 갑자기 대화가 끊겼다. 나는 <u>그녀</u> 의 골격이 그리는 근사한 선과 <u>그녀</u> 복부의 삼각형 근육을 <u>눈으로</u> 더듬었다.			
Model	P1	P2	P3	P4
GPT	(ES: -1)...나는 <u>그녀</u> 의 놀랍도록 아름다운 뼈대의 윤곽과 배에 잡힌 삼각형의 근육을...	(ES: -2)...나는 <u>멜라니</u> 의 놀랍도록 아름다운 뼈대의 윤곽과 배의 근육이 이루는 삼각형을...	(ES: -1)...나는 <u>그녀</u> 의 놀라운 뼈대의 윤곽과 배에 있는 삼각형의 근육을...	(ES: -2)...나는 <u>멜라니</u> 의 놀랍도록 아름다운 뼈대의 윤곽과 배 위에 잡힌 삼각형의 근육을...
Gemini	(ES: -1)...나는 <u>그녀</u> 의 경이로운 뼈대와 복부의 삼각형 근육 윤곽을...	(ES: -2)...나는 <u>멜라니</u> 의 놀라운 뼈대 윤곽과 복부의 삼각형 근육을...	(ES: -2)...나는 <u>멜라니</u> 의 경이로운 뼈대와 복부의 삼각형 근육 윤곽을...	(ES: -2)...나는 <u>멜라니</u> 의 경이로운 뼈대와 복부의 삼각형 근육 윤곽을...
Claude	(ES: -1)...나는 <u>그녀</u> 의 경이로운 뼈의 윤곽과 배의 삼각형 근육을...	(ES: -2)...나는 <u>멜라니</u> 의 경이로운 뼈의 윤곽과 배의 삼각형 근육을...	(ES: -1)...나는 <u>멜라니</u> 의 경이로운 뼈의 윤곽과 <u>그녀</u> 의 배에 있는 삼각형 근육을...	(ES: -2)...나는 <u>멜라니</u> 의 경이로운 뼈의 윤곽과 배 위에 새겨진 삼각형의 근육을...

Because P2 and P3 both supply the proper name, P2 is the appropriate reference for gauging the retention instruction, and compliance here entails a decrease in impicitation. Under P3, GPT and Claude both moved in that direction (OR = 0.41, $z = -3.29$, $p = .006$; OR = 0.05, $z = -7.37$, $p < .001$), although both remained well above the human baseline; Claude's reduction was by far the larger, bringing its predicted probability from .996 to .937. Gemini, by contrast, did not differ from its P2 rate (OR = 0.87, $z = -0.26$, $p = .994$), remaining at .997. Model differences realigned under this prompt: Gemini exceeded GPT (OR = 15.03, $z = 6.59$, $p < .001$), while Claude fell below Gemini (OR = 0.05, $z = -7.41$, $p < .001$) and no longer differed from GPT (OR = 0.72, $z = -1.34$, $p = .375$).

Under P4, the question was again whether a directive toward literary naturalness would shift gender handling beyond what the provision of the proper name had already

established. It did not: none of the three models differed from its P2 rate (GPT, $OR = 0.51$, $z = -2.44$, $p = .070$; Gemini, $OR = 2.13$, $z = 1.20$, $p = .630$; Claude, $OR = 0.60$, $z = -1.11$, $p = .682$), and predicted probabilities remained high throughout (.962, .999, and .994). Against P3, Claude increased significantly ($OR = 11.05$, $z = 6.99$, $p < .001$), whereas neither GPT nor Gemini did ($OR = 1.25$, $z = 0.89$, $p = .813$; $OR = 2.44$, $z = 1.44$, $p = .475$), indicating that only Claude drew a statistically reliable distinction between the retention and naturalness instructions.

Overall, the significant model $\times$ prompt interaction indicates that the effects of prompting on gender implicitation were model-dependent in the English-to-Korean data.

5. Conclusion and discussion

5.1 Conclusion

This study asked whether gender explicitation is language-dependent: whether it is paralleled by gender implicitation in the reverse direction. The results show a clear bidirectional pattern: gender information that was unmarked in Korean tended to become overt in English, whereas overt feminine marking in English tended to be reduced or omitted in Korean. The direction of the shift therefore followed target-language conventions rather than a uniform tendency toward greater explicitness.

This overall tendency was nevertheless mediated by model and prompting condition. The models differed in their default treatment of gender reference and in their responsiveness to prompting, and the provision of protagonist information exerted a more consistent influence than explicit instructions either to retain source-text gender marking or to produce a natural literary translation. Prompt effects should therefore not be understood as uniform properties of LLM translation but as conditional on how individual models respond to the information and instructions supplied.

The human translations moved in the same general direction as the LLM outputs but accommodated target-language conventions more moderately. This difference is particularly relevant in literary translation, where the overt marking or suppression of gender can affect referential ambiguity, characterization, relational disclosure, and reader interpretation even when alternative renderings remain grammatically and propositionally acceptable. Taken together, the findings suggest that gender explicitation and implicitation are best understood

as language-pair- and discourse-sensitive translation choices, and that evaluation of LLM literary translation should attend to such choices alongside conventional measures of accuracy and fluency.

5.2 Discussion

The bidirectional pattern of gender marking observed in this study can be compared with previous accounts of symmetry and asymmetry in explicitation. Klaudy and Károly's (2005) asymmetry hypothesis distinguishes between symmetric explicitation, where explicitation in the L1→L2 direction is matched by implicitation in the L2→L1 direction, and asymmetric explicitation, where no such counterpart occurs; the hypothesis holds that translators, given a choice, tend toward explicitation and often forgo optional implicitation. For gender reference in the present data, the relationship proved symmetric: the Korean-to-English direction produced gender explicitation and the English-to-Korean direction gender implicitation, each aligning with the conventions of the respective target language. What might appear as a general directional bias in translation thus resolves, in this domain, into a function of target-language norms.

In this respect, the present findings converge with the explanatory principle underlying Becher's (2010, 2011) critique of translation-inherent explicitation, although they should not be taken as a direct extension of his empirical findings from connectives to personal reference. Becher's analyses focused primarily on connectives and related discourse phenomena; their relevance to the present study lies instead in his argument that directional shifts may be explained through structural and communicative differences specific to a language pair. In the present case, the pressures operating in the two directions differ in kind. English overtly realizes third-person singular reference more frequently and encodes gender in personal pronouns, whereas Korean permits extensive argument omission and may stylistically disfavor repeated overt feminine reference. As S. W. Lee (2012) notes, explicitly gendered third-person forms such as *geu nyeo* may be perceived as unnatural in Korean and are often avoided in translation from English. The reversal observed here—greater overt gender marking in Korean-to-English translation and reduced marking in English-to-Korean translation—is therefore consistent with an account based on target-language-specific grammatical and stylistic conventions rather than a general tendency intrinsic to translation. The point of convergence with Becher is thus the

explanatory principle, rather than the identity of the linguistic phenomena examined.

Human and model behavior diverge, however, in a way the norm account alone does not capture. As Figures 2 and 3 show, the human baseline occupied an intermediate position in both directions, shifting in 46.2% of sentences in the Korean-to-English data and 67.6% in the English-to-Korean data, whereas every model exceeded that level once the protagonist's name was supplied. The direction of the human translator's choices thus followed target-language convention, as the models' did, but the degree of accommodation was consistently more moderate. One possible reading is that this restraint reflects a consideration specific to queer narrative, where marking the referent's gender also specifies the nature of the relationship; the models, by contrast, appear to respond to target-language convention alone. This interpretation remains tentative, as the present design does not isolate genre-related motivations, and research on translation norms in queer literary texts is not yet sufficiently developed to support it directly.

The prompting results also show continuity with, but not uniform replication of, earlier findings. The comparatively greater responsiveness of GPT to the gender-retention instruction in the Korean-to-English direction is consistent with patterns previously reported for GPT-4o (Yim 2025a, 2025c). In the present study, however, the same instruction did not produce a comparable response across models, reinforcing the finding that prompt effects are model-dependent rather than general properties of LLM translation.

Two patterns fell outside the study's central questions but warrant brief mention. The first concerns misgendering. Under the baseline prompt (P1), where the isolated source sentence provided no explicit basis for identifying the referent's gender, the models nonetheless assigned gender in a substantial share of cases (6.3%; Table 5). These assignments appear to track stereotypical associations rather than textual evidence: discarded cigarette butts and a coolly measured voice elicited masculine forms (Examples 1 and 2), whereas cooking elicited feminine ones (Example 3). Notably, the feminine forms in Example 3 happen to match the referent's actual gender, yet rest on the same inferential basis as the masculine ones; the accuracy of the outcome does not follow from the soundness of the inference.

Table 5. Third-person singular they/misgendering (frequency per 100)

Corpus	Third-person singular they	Misgendering
By model		
Human	0 (0.00)	0 (0.00)
CLD	93 (3.07)	83 (2.74)
GEM	155 (4.88)	56 (1.76)
GPT	90 (3.00)	94 (3.14)
AI total	338 (3.67)	233 (2.53)
By prompt		
Human	0 (0.00)	0 (0.00)
P1	321 (14.70)	138 (6.32)
P2	1 (0.04)	30 (1.28)
P3	7 (0.30)	14 (0.61)
P4	9 (0.38)	51 (2.16)
AI total	338 (3.67)	233 (2.53)

(1) Cases of misgendering 1

ST: 그 애의 발이 아무렇게나 버려진 담배꽂초들을 하나씩 터뜨린다.

HT: Lane crushes and bursts the cigarette butts strewn on the ground.

K173_GPT_P1_run1: His feet pop discarded cigarette butts one by one.

(2) Cases of misgendering 2

ST: 딸애의 목소리는 뜨겁고 그 애의 목소리는 적당히 서늘하다.

HT: My daughter's voice is hot and Lane's voice is just cool enough.

K035_Gemini_P1: My daughter's voice is hot, and his voice is moderately cool.

K035_Claude_P1: My daughter's voice is hot and his voice is pleasantly cool.

(3) Cases of misgendering 3

ST: 나는 그 애가 만들었을 게 분명한 음식들을 맛본다.

'I taste the dishes that must have been made by that child.'

HT: I taste the food that I'm sure Lane made.

K062_GPT-P1: I taste the dishes that the child must have made.

K062_Gemini_P1: I taste the food that she must have made.

K062_Claude_P1: I taste the food that she must have made.

The frequency of misgendering declined once the protagonist's name was supplied

(1.3% under P2), consistent with reports that contextual information mitigates gender bias in LLM output (Hendy et al. 2023; Dawkins et al. 2024; Savoldi et al. 2024; Tang et al. 2025). It did not disappear under the prompts that supplied the name, however, rising again to 2.2% under P4, where the instruction to translate naturally appears to have loosened that constraint. Contextual disambiguation thus limits the underlying tendency without eliminating it, and bias in this domain merits continued investigation.

(4) Cases of third-person singular *they*

ST: 그 애는 먹던 빵을 내려놓고 가만히 입가를 닦는다.

HT: She puts down the sandwich and carefully dabs at the corners of her mouth.

K087_Gemini_P1_run2: The child puts down the bread they were eating and quietly wipes their mouth.

The second concerns singular *they*. Its use was almost entirely confined to the baseline prompt (14.7%, against 0.0-0.4% elsewhere; Table 5), indicating that the models treated it as a device for managing referential uncertainty rather than as a general option, as shown in Example 4. Even there it remained a minority strategy, with explicit gender marking the more common response to an unmarked source. This is consistent with prior findings on the difficulty LLMs have in deploying singular *they* consistently (Hossain et al. 2023; Friðriksdóttir 2024; Yim 2025a), and may also reflect the limited establishment of singular *they* in literary prose, notwithstanding its increasing endorsement in style guidance. Building on proposals for gender-neutral language practice (Krauthamer 2021), future evaluation frameworks might incorporate gender-neutral rendering as an explicit criterion.

Finally, the study draws on a single literary work in each direction, and its findings should be generalized with caution: the patterns observed here may reflect properties of these particular texts as much as of the language pair, although the two were selected for comparability in theme and narrative situation. Another limitation stems from the sentence-level design. Because surrounding discourse was withheld for experimental control, the study does not assess gender resolution under fuller narrative context. Future research could compare these sentence-level patterns with passage- or document-level translation. A further limitation concerns interpretability: because

the models' internal processing is not observable through the API, the design shows that a given prompt altered the pattern of overt gender marking, but not why it did so. The findings are therefore interpreted as observed patterns under controlled prompting conditions, not as evidence of how the models arrive at them.

The narrow scope nonetheless provides a starting point for further bidirectional research on gender rendering in English–Korean translation. The findings also reveal an important difference between model and human translation: although both shifted gender marking toward target-language conventions, the human translations did so more moderately, whereas the models tended to accommodate those conventions much more uniformly once contextual information was available. This difference is consequential because overt gender marking is not merely a question of grammatical well-formedness or propositional accuracy. Where multiple linguistically acceptable renderings are possible, the decision to overtly mark or suppress gender can constitute a point of translational intervention, potentially affecting referential ambiguity, characterization, reader interpretation, and related literary effects. Evaluation of LLMs for literary and other creative texts therefore needs to account for such discourse-sensitive choices in addition to sentence-level accuracy.

References

- Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers (ed.), *Terminology, LSP and translation*, 175–186. Amsterdam: John Benjamins.
- Bas, Tetiana. 2024. Assessing gender bias in LLMs: Comparing LLM outputs with human perceptions and official statistics. *arXiv preprint arXiv:2411.13738*. <https://doi.org/10.48550/arXiv.2411.13738>.
- Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1): 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Becher, Viktor. 2010. Abandoning the notion of "translation-inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures* 11(1): 1–28. <https://doi.org/10.1556/Acr.11.2010.1.1>.
- Becher, Viktor. 2011. When and why do translators add connectives? A corpus-based study. *Target* 23(1): 26–47. <https://doi.org/10.1075/target.23.1.02bec>.

- Cho, Won Ik, Jiwon Kim, Jaeyeong Yang, and Nam Soo Kim. 2021. Towards cross-lingual generalization of translation gender bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 449–457. New York, NY: Association for Computing Machinery. <https://dl.acm.org/doi/10.1145/3442188.3445907>.
- Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.
- Curzan, Anne. 2003. *Gender shifts in the history of English*. Cambridge: Cambridge University Press.
- Dawkins, Hillary, Isar Nejadgholi, and Chi-Kiu Lo. 2024. WMT24 test suite: Gender resolution in speaker–listener dialogue roles. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, 307–326. Stroudsburg, PA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.wmt-1.25>.
- Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5): 378–382. <https://doi.org/10.1037/h0031619>.
- von Flotow, Luise. 1997. *Translation and gender: Translating in the “era of feminism”*. Manchester: St. Jerome.
- Friðriksdóttir, Steinunn Rut. 2024. The GenderQueer test suite. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, 327–340. Stroudsburg, PA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.wmt-1.26>.
- Gao, Ruiyao, Yumeng Lin, Nan Zhao, and Zhenguang G. Cai. 2024. Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications* 11(1): 835. <https://doi.org/10.1057/s41599-024-03363-0>.
- Goh, Yong-Geun and Koo Bon-Kwan. 2018. *Urilmal munbeopron [Korean grammar theory]*. Seoul: Jipmundang.
- Guan, Xingzhong. 2024. A bard is born: A new era of poetry translation by ChatGPT-4. *Translation Review* 120(1): 23–39. <https://doi.org/10.1080/07374836.2024.2365778>.
- Gwet, Kilem Li. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology* 61(1): 29–48. <https://doi.org/10.1348/000711006X126600>.
- Han, Na-Rae. 2006. *Korean zero pronouns: Analysis and resolution*. PhD Dissertation. University of Pennsylvania.
- He, Sui. 2024. Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. *arXiv preprint arXiv:2403.00127*. <https://doi.org/10.48550/arXiv.2403.00127>.

- Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*. <https://doi.org/10.48550/arXiv.2302.09210>.
- Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied logistic regression*. 3rd ed. Hoboken, NJ: Wiley. <https://doi.org/10.1002/9781118548387>.
- Hossain, Tamanna, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. *arXiv preprint arXiv:2306.03950*. <https://doi.org/10.48550/arXiv.2306.03950>.
- Irshad, Isra and Musarat Yasmin. 2022. Feminism and literary translation: A systematic review. *Heliyon* 8(3): e09082. <https://doi.org/10.1016/j.heliyon.2022.e09082>.
- Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*. <http://arxiv.org/abs/2301.08745>.
- Karaban, Viacheslav and Anna Karaban. 2024. AI-translated poetry: Ivan Franko's poems in GPT-3.5-driven machine and human-produced translations. *Forum for Linguistic Studies* 6(1): 1994. <https://doi.org/10.59400/fls.v6i1.1994>.
- Kim, Han-sik, Dong-hee Kang, Seul-gi Nam, Seung-hee Seo, Joo-hee Seok, Shin-ae Song, Ji-soo Choi, and Seung-bin Hong. 2019. A-B translations of null subjects: Focusing on Plaza Hotel. *Interpretation and Translation* 21(3): 31–54. <https://doi.org/10.20305/it201903031054>.
- Kim, Hye-jin. 2017. *Ttal-e daeha-yeo [Concerning my daughter]*. Seoul: Minumsa.
- Kim, Hye-jin. 2022. *Concerning my daughter*. Translated by Jamie Chang. London: Picador.
- Klaudy, Kinga. 2009. Explication. In Mona Baker and Gabriela Saldanha (eds.), *Encyclopaedia of translation studies*, 2nd edition, 104–108. London: Routledge.
- Klaudy, Kinga and Krisztina Károly. 2005. Implication in translation: Empirical evidence for operational asymmetry in translation. *Across Languages and Cultures* 6(1): 13–28. <https://doi.org/10.1556/acr.6.2005.1.2>.
- Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinhórfur Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, 1–46. Stroudsburg, PA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.wmt-1.1>.
- Krauthamer, Helene S. 2021. *The great pronoun shift: The big impact of little parts of speech*. 1st ed. London: Routledge. <https://doi.org/10.4324/9780367210175>.

- Lardelli, Manuel. 2024. Gender-fair translation: A case study beyond the binary. *Perspectives* 32(6): 1146–1162. <https://doi.org/10.1080/0907676X.2023.2268654>.
- Lee, Sang-Bin. 2013. A reader-response questionnaire survey on the Korean pronoun *geunyeo*: A comparison of university students and elementary school students. *Interpreting and Translation Studies* 17(4): 121–137.
- Lee, Sang Won. 2012. Korean pronoun 'keunyo' and translation. *Interpretation and Translation* 14(1): 193–211.
- Liang, Weixin, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. 2025. The widespread adoption of large language model-assisted writing across society. *Patterns* 6(12): 101366. <https://doi.org/10.1016/j.patter.2025.101366>
- Miletich, Marko. 2012. *Reading gender in translation: Translator's intervention in Isaac Chocrón's "Pronombres personales."* PhD Dissertation. State University of New York at Binghamton.
- Ngo, Hong Hai, Yunmeng Li, and Keisuke Sakaguchi. 2025. Towards equitable translation: Gender bias in large language models. In Association for Natural Language Processing (ed.), *Proceedings of the Thirty-first Annual Meeting of the Association for Natural Language Processing*, 3894–3898. Kyoto: Association for Natural Language Processing. https://www.anlp.jp/proceedings/annual_meeting/2025/pdf_dir/A10-3.pdf.
- Piazzolla, Silvia Alma, Beatrice Savoldi, and Luisa Bentivogli. 2023. 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/hjlc.vi63.137553>.
- Rose, Emily. 2018. Surmounting the “insurmountable” challenges of translating a transgender memoir. In Duncan Large, Motoko Akashi, Wanda Józwickowska, and Emily Rose (eds.), *Untranslatability: Interdisciplinary perspectives*, 161–178. London: Routledge.
- Rossi, María Julia. 2021. Of epicene particles and other misleading tricks: Gender ambiguity in Silvina Ocampo’s “Carta perdida en un cajón.” *Textual Cultures* 14(1): 189–207. <https://doi.org/10.14434/tc.v14i1.32857>.
- Sant, Aleix, 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. *arXiv preprint arXiv:2407.18786*.
- Savoldi, Beatrice, Andrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024. A prompt response to the demand for automatic gender-neutral translation. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, 256–267. Stroudsburg, PA: Association for Computational Linguistics.
- Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics* 9: 845–874. https://doi.org/10.1162/tac1_a_00401.

- Song, Sooho. 2022. Analysis of gender pronoun errors in Korean speakers' English speech. *English Teaching* 77(1): 21–39. <https://doi.org/10.15858/engtea.77.1.202203.21>.
- Tang, Xushuo, Yi Ding, Zhengyi Yang, Yin Chen, Yongrui Gu, Wenke Yang, Mingchen Ju, Xin Cao, Yongfei Liu, and Wenjie Zhang. 2025. Do they understand them? An updated evaluation on nonbinary pronoun handling in large language models. *arXiv preprint arXiv:2508.00788*. <https://doi.org/10.48550/arXiv.2508.00788>.
- Vanmassenhove, Eva. 2024. Gender bias in machine translation and the era of large language models. *arXiv preprint arXiv:2401.10016*. <https://doi.org/10.48550/arXiv.2401.10016>.
- Winterson, Jeanette. 1997. *Oranges are not the only fruit*. New York: Grove Atlantic.
- Winterson, Jeanette. 2024. *Orenjimani gwaireun anida* [Oranges are not the only fruit]. Translation by Kim Eun-jeong. Seoul: Minumsa.
- Yim, Jin. 2025a. Gender ambiguity in human and AI translations: “That person” in Concerning my daughter. *Korean Journal of English Language and Linguistics* 25: 912–934. <https://doi.org/10.15738/kjell.25..202507.912>.
- Yim, Jin. 2025b. AI-based machine translation of gender-neutral expressions: Focusing on literary translations of GPT-4o, Gemini, and DeepL. *Interpreting and Translation Studies* 29(3): 61–84. <https://doi.org/10.22844/its.2025.29.3.61>.
- Yim, Jin. 2025c. Preserving ambiguity: Prompt sensitivity in gender-neutral literary translation by GPT models. In Chu-Ren Huang, Yasunari Harada, Jong-Bok Kim, Nguyen T.M. Huyen, Le Thanh Huong, Pham Hien, Emmanuele Chersoni, Le Minh Nguyen, Rachel Edita Oñate Roxas, and Sherly Dita (eds.), *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, 113–125. Hanoi, Vietnam: Association for Computational Linguistics.
- Zhang, Ran, Wei Zhao, and Steffen Eger. 2024. How good are LLMs for literary translation, really? Literary translation evaluation with humans and LLMs. *arXiv preprint arXiv:2410.18697*. <https://doi.org/10.48550/arXiv.2410.18697>.
- Zhou, Pei, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5276–5284. Hong Kong, China: Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1531>.

Appendix 1. Prompts

Korean to English Translation	
P1	Translate the following Korean sentence into English:
P2	Translate the following Korean sentence into English. In this text, the character is a woman named Lane.
P3	Translate the following Korean sentence into English. In this text, the character is a woman named Lane. Preserve the same degree of gender marking as the source text.
P4	Translate the following Korean sentence into English. In this text, the character is a woman named Lane. Translate as natural literary prose.
English to Korean Translation	
P1	Translate the following English sentence into Korean:
P2	Translate the following English sentence into Korean. In this text, her name is Melanie.
P3	Translate the following English sentence into Korean. In this text, her name is Melanie. Preserve the same degree of gender marking as the source text.
P4	Translate the following English sentence into Korean. In this text, her name is Melanie. Translate as natural literary prose.
Common	
Return ONLY the translation. Do not explain. Do not comment. Do not output the source sentence. Do not output the prompt. Do not mention the instructions. Do not mention the character information. Do not use quotation marks. Output exactly one translated sentence.	

Appendix 2. Basic corpus information (KE)

Model	Prompt	Token	Token (/sent.)	ES (sum)	ES (/sent.)	ES>0	ES=0
CLD	P1	2773	16.4	80	0.47	55	114
	P1	2756	16.3	79	0.47	55	114
	P1	2766	16.4	81	0.48	57	112
	P2	2634	15.6	198	1.17	141	28
	P2	2628	15.6	193	1.14	141	28
	P2	2645	15.7	198	1.17	144	25
	P3	2614	15.5	212	1.25	153	16
	P3	2597	15.4	207	1.22	150	19

	P3	2598	15.4	207	1.22	152	17
	P4	2730	16.2	204	1.21	142	27
	P4	2738	16.2	209	1.24	142	27
	P4	2752	16.3	209	1.24	143	26
GEM	P1	2657	15.7	113	0.67	83	86
	P1	2652	15.7	116	0.69	86	83
	P1	2658	15.7	114	0.67	83	86
	P2	2587	15.3	109	0.65	81	88
	P2	2583	15.3	106	0.63	77	92
	P2	2589	15.3	109	0.65	82	87
	P3	2582	15.3	156	0.92	118	51
	P3	2578	15.3	150	0.89	113	56
	P3	2594	15.3	157	0.93	117	52
	P4	2648	15.7	125	0.74	91	78
	P4	2630	15.6	115	0.68	87	82
	P4	2610	15.4	117	0.69	85	84
GPT	P1	2754	16.3	27	0.16	22	147
	P1	2751	16.3	28	0.17	22	147
	P1	2739	16.2	27	0.16	21	148
	P2	2670	15.8	135	0.8	99	70
	P2	2678	15.8	146	0.86	107	62
	P2	2664	15.8	138	0.82	96	73
	P3	2748	16.3	89	0.53	63	106
	P3	2763	16.3	99	0.59	66	103
	P3	2756	16.3	96	0.57	69	100
	P4	2676	15.8	121	0.72	89	80
	P4	2699	16	128	0.76	90	79
	P4	2722	16.1	127	0.75	90	79
Human	N/A	2373	14	99	0.59	78	91

Appendix 3. Translation choices in KE corpus

Panel A: By model

Category	Human	CLD	GEM	GPT	AI total
Gender pronoun	85 (39.0%)	1972 (65.2%)	1525 (48.0%)	1087 (36.3%)	4584 (49.8%)
Gender noun	18 (8.3%)	228 (7.5%)	72 (2.3%)	160 (5.3%)	460 (5.0%)
Neutral pronoun	7 (3.2%)	263 (8.7%)	321 (10.1%)	275 (9.2%)	859 (9.3%)
Neutral noun	2 (0.9%)	426 (14.1%)	249 (7.8%)	1056 (35.2%)	1731 (18.8%)

Proper name	106 (48.6%)	137 (4.5%)	1009 (31.8%)	418 (14.0%)	1564 (17.0%)
Total expressions	218	3026	3176	2996	9198

Panel B: By prompt

Category	Human	P1	P2	P3	P4	AI total
Gender pronoun	85 (39.0%)	698 (32.0%)	1295 (55.2%)	1345 (58.2%)	1246 (52.9%)	4584 (49.8%)
Gender noun	18 (8.3%)	21 (1.0%)	127 (5.4%)	111 (4.8%)	201 (8.5%)	460 (5.0%)
Neutral pronoun	7 (3.2%)	476 (21.8%)	131 (5.6%)	133 (5.8%)	119 (5.1%)	859 (9.3%)
Neutral noun	2 (0.9%)	988 (45.3%)	175 (7.5%)	388 (16.8%)	180 (7.6%)	1731 (18.8%)
Proper name	106 (48.6%)	0 (0.0%)	619 (26.4%)	335 (14.5%)	610 (25.9%)	1564 (17.0%)
Total expressions	218	2183	2347	2312	2356	9198

The frequencies reported in the tables represent the simple sum of occurrences across all translation runs.

Appendix 4. Translation choices in EK corpus

Panel A: By model

Category	Human	CLD	GEM	GPT	AI total
Gender pronoun	53 (47.3%)	647 (41.9%)	521 (34.5%)	689 (45.1%)	1857 (40.5%)
Gender noun	0 (0.0%)	0 (0.0%)	3 (0.2%)	1 (0.1%)	4 (0.1%)
Neutral pronoun	2 (1.8%)	51 (3.3%)	33 (2.2%)	38 (2.5%)	122 (2.7%)
Neutral noun	0 (0.0%)	0 (0.0%)	1 (0.1%)	0 (0.0%)	1 (0.0%)
Proper name	57 (50.9%)	845 (54.8%)	954 (63.1%)	800 (52.4%)	2599 (56.7%)
Total expressions	112	1543	1512	1528	4583

Panel B: By prompt

Category	Human	P1	P2	P3	P4	AI total
Gender pronoun	53 (47.3%)	1045 (92.1%)	231 (19.6%)	345 (29.7%)	236 (21.3%)	1857 (40.5%)
Gender noun	0 (0.0%)	3 (0.3%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	4 (0.1%)

Neutral pronoun	2 (1.8%)	32 (2.8%)	36 (3.1%)	27 (2.3%)	27 (2.4%)	122 (2.7%)
Neutral noun	0 (0.0%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)
Proper name	57 (50.9%)	54 (4.8%)	908 (77.2%)	790 (68.0%)	847 (76.3%)	2599 (56.7%)
Total expressions	112	1135	1176	1162	1110	4583

The frequencies reported in the tables represent the simple sum of occurrences across all translation runs.

Appendix 5. Basic corpus information (EK)

Model	Prompt	Token	Token (/sent.)	ES (sum)	ES (/sent.)	ES<0	ES=0
CLD	P1	1066	10.451	-31	-0.30	22	80
	P1	1071	10.500	-31	-0.30	22	80
	P1	1074	10.529	-31	-0.30	22	80
	P2	1082	10.608	-123	-1.21	98	3
	P2	1079	10.578	-122	-1.20	98	3
	P2	1076	10.549	-121	-1.19	96	5
	P3	1053	10.324	-97	-0.95	78	24
	P3	1063	10.422	-95	-0.93	76	26
	P3	1056	10.353	-96	-0.94	77	25
	P4	1089	10.676	-121	-1.19	95	6
	P4	1079	10.578	-124	-1.22	96	5
GEM	P4	1079	10.578	-124	-1.22	96	5
	P1	1054	10.333	-35	-0.34	25	77
	P1	1053	10.324	-35	-0.34	25	77
	P1	1059	10.382	-35	-0.34	25	77
	P2	1051	10.304	-125	-1.23	98	2
	P2	1049	10.284	-126	-1.24	99	2
	P2	1052	10.314	-124	-1.22	97	4
	P3	1055	10.343	-124	-1.22	98	4
	P3	1057	10.363	-123	-1.21	97	5
	P3	1051	10.304	-124	-1.22	98	4
	P4	1067	10.461	-129	-1.26	99	2
	P4	1062	10.412	-129	-1.26	100	2
GPT	P4	1079	10.578	-128	-1.25	99	2
	P1	1086	10.647	-32	-0.31	22	80
	P1	1099	10.775	-31	-0.30	22	80
	P1	1098	10.765	-32	-0.31	22	80

	P2	1110	10.882	-115	-1.13	92	8
	P2	1117	10.951	-114	-1.12	89	10
	P2	1113	10.912	-112	-1.10	86	14
	P3	1103	10.814	-103	-1.01	81	19
	P3	1113	10.912	-101	-0.99	79	22
	P3	1105	10.833	-108	-1.06	82	20
	P4	1145	11.225	-113	-1.11	88	13
	P4	1143	11.206	-104	-1.02	79	21
	P4	1137	11.147	-109	-1.07	82	19
Human	N/A	1032	10.118	-93	-0.91	69	32

Appendix 6. Tukey-adjusted pairwise comparisons among models

Prompt	Contrast	Odds ratio	SE	95% CI	z	p
Korean-to-English Translation						
P1	GEM / GPT	11.7650	2.1600	[7.6520, 18.0880]	13.432	< .001
	CLD / GPT	4.6660	0.8630	[3.0240, 7.1980]	8.326	< .001
	CLD / GEM	0.3970	0.0591	[0.2800, 0.5620]	-6.203	< .001
P2	GEM / GPT	0.5230	0.0758	[0.3720, 0.7350]	-4.470	< .001
	CLD / GPT	4.8220	0.8080	[3.2550, 7.1420]	9.385	< .001
	CLD / GEM	9.2190	1.5500	[6.2140, 13.6760]	13.198	< .001
P3	GEM / GPT	4.9700	0.7520	[3.4860, 7.0840]	10.598	< .001
	CLD / GPT	25.8480	4.9400	[16.5200, 40.4440]	17.027	< .001
	CLD / GEM	5.2010	0.9860	[3.3350, 8.1120]	8.696	< .001
P4	GEM / GPT	0.9400	0.1350	[0.6710, 1.3160]	-0.434	.901
	CLD / GPT	6.9340	1.1600	[4.6770, 10.2800]	11.526	< .001
	CLD / GEM	7.3810	1.2400	[4.9770, 10.9450]	11.891	< .001
English-to-Korean Translation						
P1	GEM / GPT	1.8526	0.6956	[0.7684, 4.4663]	1.642	.228
	CLD / GPT	1.0000	0.3875	[0.4032, 2.4800]	0.000	1.000
	CLD / GEM	0.5398	0.2027	[0.2239, 1.3015]	-1.642	.228
P2	GEM / GPT	7.0199	3.0923	[2.5001, 19.7108]	4.424	< .001
	CLD / GPT	5.3776	2.2004	[2.0612, 14.0302]	4.111	< .001
	CLD / GEM	0.7661	0.3968	[0.2275, 2.5794]	-0.514	.864
P3	GEM / GPT	15.0330	6.1861	[5.7306, 39.4361]	6.586	< .001
	CLD / GPT	0.7216	0.1762	[0.4072, 1.2787]	-1.337	.375
	CLD / GEM	0.0480	0.0197	[0.0184, 0.1255]	-7.408	< .001
P4	GEM / GPT	29.3440	16.0001	[8.1756, 105.3213]	6.197	< .001
	CLD / GPT	6.3743	2.2193	[2.8187, 14.4150]	5.320	< .001
	CLD / GEM	0.2172	0.1271	[0.0551, 0.8563]	-2.609	.025

Pairwise comparisons were adjusted using Tukey's method within each prompt. Odds ratios compare the first with the second model in each contrast; a value below 1 indicates that the first had lower odds of gender explicitation (in Korean-to-English translation) or implication (in English-to-Korean translation) than the second. Estimates were averaged over run.

Appendix 7. Tukey-adjusted pairwise comparisons among prompts

Model	Contrast	Odds ratio	SE	95% CI	z	p
Korean-to-English Translation						
GPT	P2 / P1	19.8460	3.6800	[12.3240, 31.9600]	16.110	< .001
GPT	P3 / P1	6.6300	1.2200	[4.1370, 10.6270]	10.301	< .001
GPT	P3 / P2	0.3340	0.0492	[0.2290, 0.4880]	-7.443	< .001
GPT	P4 / P1	14.0400	2.5800	[8.7520, 22.5240]	14.361	< .001
GPT	P4 / P2	0.7070	0.1020	[0.4880, 1.0260]	-2.394	.078
GPT	P4 / P3	2.1180	0.3090	[1.4550, 3.0820]	5.135	< .001
GEM	P2 / P1	0.8820	0.1270	[0.6100, 1.2770]	-0.870	.821
GEM	P3 / P1	2.8010	0.4160	[1.9130, 4.1010]	6.939	< .001
GEM	P3 / P2	3.1740	0.4720	[2.1660, 4.6520]	7.764	< .001
GEM	P4 / P1	1.1210	0.1610	[0.7750, 1.6220]	0.796	.856
GEM	P4 / P2	1.2710	0.1830	[0.8780, 1.8390]	1.665	.343
GEM	P4 / P3	0.4000	0.0593	[0.2740, 0.5860]	-6.175	< .001
CLD	P2 / P1	20.5090	3.5700	[13.1120, 32.0790]	17.349	< .001
CLD	P3 / P1	36.7310	7.1300	[22.3140, 60.4620]	18.575	< .001
CLD	P3 / P2	1.7910	0.3640	[1.0630, 3.0180]	2.868	.022
CLD	P4 / P1	20.8660	3.6400	[13.3250, 32.6770]	17.402	< .001
CLD	P4 / P2	1.0170	0.1890	[0.6310, 1.6390]	0.093	> .999
CLD	P4 / P3	0.5680	0.1160	[0.3370, 0.9580]	-2.778	.028
English-to-Korean Translation						
GPT	P2 / P1	901.6170	396.6720	[291.1761, 2791.8300]	15.466	< .001
GPT	P3 / P1	366.1960	153.2090	[125.0027, 1072.7700]	14.110	< .001
GPT	P3 / P2	0.4060	0.1110	[0.2008, 0.8200]	-3.287	.006
GPT	P4 / P1	458.2430	193.8330	[154.5814, 1358.4200]	14.486	< .001
GPT	P4 / P2	0.5080	0.1410	[0.2491, 1.0400]	-2.438	.070
GPT	P4 / P3	1.2510	0.3170	[0.6526, 2.4000]	0.885	.813
GEM	P2 / P1	3416.4250	1790.8580	[888.6361, 13134.6900]	15.522	< .001
GEM	P3 / P1	2971.5390	1514.2530	[802.4695, 11003.5900]	15.693	< .001
GEM	P3 / P2	0.8700	0.4600	[0.2236, 3.3800]	-0.264	.994
GEM	P4 / P1	7258.3050	4527.3380	[1461.8805, 36037.8300]	14.252	< .001
GEM	P4 / P2	2.1250	1.3390	[0.4207, 10.7300]	1.195	.630
GEM	P4 / P3	2.4430	1.5150	[0.4962, 12.0200]	1.440	.475
CLD	P2 / P1	4848.5330	2569.4600	[1242.6283, 18918.1900]	16.014	< .001
CLD	P3 / P1	264.2400	108.9480	[91.6175, 762.1100]	13.526	< .001
CLD	P3 / P2	0.0540	0.0220	[0.0198, 0.1500]	-7.369	< .001
CLD	P4 / P1	2920.9590	1436.5770	[825.6526, 10333.6500]	16.225	< .001
CLD	P4 / P2	0.6020	0.2740	[0.1869, 1.9400]	-1.112	.682
CLD	P4 / P3	11.0540	3.8000	[4.5711, 26.7300]	6.990	< .001

Pairwise comparisons were adjusted using Tukey's method within each model. Odds ratios compare the first with the second prompt in each contrast; a value below 1 indicates that the first had lower odds of gender explicitation (in Korean-to-English translation) or implicitation (in English-to-Korean translation) than the second. Estimates were averaged over run.

Jin Yim

Adjunct Lecturer

Graduate School of Translation and Interpretation

Ewha Womans University

52 Ewhayeodae-gil, Seodaemun-gu,

Seoul, 03760, Korea

E-mail: jy2812@gmail.com

Received: 2026. 07. 24.

Revised: 2026. 09. 08.

Accepted: 2026. 09. 12.