"Not Polite but Not Rude." LLM Formality Adaptation Across Gendered Linguistic Styles in Japanese

Momo Takamatsu[★1] **Filip Miletić**[♦1*] **Agnieszka Falenska**[♦1,2*]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]Interchange Forum for Reflecting on Intelligent Systems, University of Stuttgart, Germany
[♦]name.surname@ims.uni-stuttgart.de [★]momo.takamatsu@googlemail.com

Abstract

Do language models adapt their linguistic style in an unbiased and socially appropriate way? We examine this issue in Japanese, a language with complex sociolinguistic constraints on stylistic adaptation between interlocutors, and we focus on the interplay between politeness and gendered linguistic styles. In terms of phenomena, we analyze morphologically encoded levels of formality alongside gender-indexical cues. In terms of methods, we analyze model behavior as well as human perceptions of generated outputs. Concretely, we first curate a novel set of 900 topic-balanced prompts with systematic formality and gendered-style variants. We then use these prompts to examine the behavior of four models, finding weak-to-moderate but consistent formality adaptation, with stronger adaptation to masculine-style than to feminine-style questions. Finally, we conduct a human perception study, showing that formality-mismatched responses are perceived as less appropriate than matched responses, with judgments strongly shaped by situational context. Taken together, our findings highlight the value of stylistic adaptation that is broadly invariant to gender-indexical cues and sensitive to situational context, with direct implications for conversational system design.

1 Introduction

Politeness in Japanese is deeply tied to cultural norms and social expectations. Linguistically, it is mainly expressed through different levels of *formality*, which are not mere stylistic devices; they convey respect, hierarchy, and interpersonal relationships (Yin et al., 2024). While similar sociolinguistic practices are found across human languages, Japanese is distinguished by a formality system that is particularly elaborate and, moreover, grammatically encoded in its morphology



Figure 1: Two questions in *polite* style, with a formality-adapting *(polite)* and non-adapting *(casual)* LLM response. Blue highlights formality-indicating sentence endings and orange gender-associated pronouns.

(Hasegawa, 2015). The core of this system, known as *keigo* (敬語; honorific speech), comprises multiple levels of formality that are selected based on interlocutors' hierarchical positions and situational context, e.g., in school or at work (Khasanah and Andari, 2025). Incorrect formality levels may be interpreted as rude or disrespectful and thus hinder communicative effectiveness by giving rise to misunderstandings and social friction (Sato, 2024).

Making matters even more complex, politeness is shaped not only by social hierarchy and situational context but also by *gender*. In Japanese society, conventional gender norms create a strong expectation that women's speech should be more polite and include more honorific markers than men's speech (Okamoto, 2013; Ide and Yoshida, 2017; Steinborn et al., 2023). This expectation further entails that deviations from polite forms are subject to stronger social judgment for women than for men. These social norms have mainly been studied in a binary framing of gender, leaving the question of how non-binary speakers navigate gendered linguistic styles across levels of formality largely underexplored (Dale, 2012; Nakamura, 2025).

*Equal contribution.

Proceedings of the 7th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 92–110
November 9, 2026 ©2026 Association for Computational Linguistics

With the growing integration of conversational large language models (LLMs) into daily life, the sociolinguistic practices discussed thus far raise important implications for human–LLM interactions (Maitra et al., 2025). An additional source of complexity is that Japanese morphologically encodes gender information not only in the formality system but also in other linguistic structures, such as personal pronouns and sentence-final particles. To illustrate the potential consequences, consider the two interactions in Figure 1. In the top interaction, the model adapts to the user's polite question and answers at the same level of formality. In the bottom, it responds less formally, potentially sounding overly familiar or disrespectful. Crucially, the questions differ only in the pronoun chosen by the user to addresses the model: the masculine-style variant uses 君 (*kimi*), whereas the feminine-style variant uses あなた (*anata*).[1] *Systematic* differences across matched prompts would indicate *gender bias*, i.e., gender-conditioned adaptation despite equivalent semantic content. Such bias can give rise to interactional and representational harms (Blodgett et al., 2020): users whose linguistic style incorporates *gender-indexical cues* – forms which index (point to) a specific gender – may be more likely to receive responses perceived as disrespectful or socially inappropriate. Moreover, systematic differences in response formality may reinforce stereotypical associations between gender, politeness, hierarchy, and social roles.

Despite these implications, formality adaptation in Japanese human–LLM interactions remains underexplored, especially regarding its interplay with gendered language and the associated sociolinguistic expectations. In order to comprehensively analyze this problem, we approach it from two perspectives – model behavior and human perception – and address the following research questions:

RQ1 Do models adapt their formality level in Japanese to the user's question, and how does this behavior vary across model families and sizes?

RQ2 Do models exhibit differences in formality adaptation based on gender-indexing linguistic styles?

RQ3 Are model responses to feminine-style questions perceived as socially appropriate, particularly when question and answer formality differ?

[1]While both 君 and あなた translate to English "you", the choice between them conveys gender-related information about the *speaker*, not the addressee.

To answer these questions, we construct a controlled benchmark of user questions that systematically vary in formality and gender-indexing linguistic style. We evaluate four LLMs from two model families, measuring their overall formality adaptation and comparing it across gender-indexing styles (§3). Our results show weak-to-moderate but consistent adaptation, with stronger effects in larger models (§4). Notably, formality adaptation is systematically stronger for masculine-style than for feminine-style questions (§5). Motivated by this asymmetry, we conduct a human perception study in which participants assess the appropriateness of formality-matched and mismatched responses to feminine-style questions. The study shows that mismatched responses are perceived as less appropriate, with judgments strongly shaped by situational context: some mismatches evoke unintended hierarchical or gendered relations and are perceived as condescending or demeaning (§6).

Our work makes three contributions: a benchmark for studying LLM formality adaptation in Japanese; a behavioral analysis of how this adaptation varies across models and gender-indexical linguistic cues; and a human perception study of its social appropriateness. Together, our findings reaffirm that formality adaptation in Japanese is not merely a matter of stylistic matching, but a socially situated behavior underpinned by gender-indexing language and broader expectations of interpersonal appropriateness.[2]

2 Related Work

We now provide further linguistic background on the expression of formality and its interaction with gender in Japanese. We then review related work on LLM robustness to language variation in general and in the more specific context of Japanese.

Formality and gender in Japanese. The Japanese honorific system *keigo* encompasses multiple levels of formality. These are primarily *sonkeigo* (尊敬語; respectful language), *kenjougo* (謙譲語; humble language), and *teineigo* (丁寧語; polite language). For instance, *sonkeigo* and *kenjougo* are typically used with authorities and in very formal settings, while *teineigo* is used in everyday interactions. In contrast, more intimate relationships, such as with friends and family, typically rely on *futsutai* (普通体; casual

[2]Our code and data are available at: https://github.com/momot12/Jp_LLM_Adapt/.

language), which stands outside the honorific system. The various formality levels can also be grouped into three coarser categories: the formal extreme (*sonkeigo* and *kenjougo*); the middle ground (*teineigo*); and the casual extreme (*futsutai*) (Marrese-Taylor et al., 2023). Importantly, the formality level is encoded in the morphological system, commonly via verbal suffixes. For example, the casual (*futsutai*) verb form *tabe**ru*** (食べる; to eat) becomes the polite (*teineigo*) form *tabe**masu*** (食べます) by changing the suffix.

The choice of formality levels in Japanese is also influenced by gender-related norms and expectations. For example, if a woman uses the verb "to eat" in the more masculine or even vulgar form *kuu* (食う), she may be perceived as unmannered, while the same usage by a man is often judged less harshly (Nakamura, 2014). Beyond these formality-related mechanisms, gender is also indexed through other linguistic cues in Japanese, such as sentence-ending particles and personal pronouns (Okamoto and Morimoto, 2023; Nakamura, 2025). When these gender-indexical cues diverge from the expectations externally associated with the speaker (e.g., given their perceived gender), the resulting linguistic style may be viewed as gender-atypical.

LLM sensitivity to language variation. LLM behavior regarding social factors of language use is increasingly studied, though mainly in English. Models adapt their outputs to different English varieties, but perform more naturally for standardized (e.g., Standard American and British English) than minority varieties, with responses to the latter more likely to be perceived as stereotypical, unnatural, or condescending (Fleisig et al., 2024). Beyond stylistic differences, variety-specific inputs can lead LLMs to reproduce harmful social biases, e.g., assigning more negative personality traits, as shown for African American English (Hofmann et al., 2024) and German dialects (Bui et al., 2025). Further, LLMs stylistically converge to human input at a higher rate than the one observed in human–human communication (Blevins et al., 2026; Chen et al., 2026), raising the prospect of unnatural interaction patterns. This effect has indeed been observed from the user perspective: language adaptation can increase social presence and engagement, but adaptations to minority language varieties may be perceived as forced or exaggerated (Basoah et al., 2025; Sandoval et al., 2025) and are more generally subject to highly variable personal preferences (Lucy et al., 2024). Overall, sociolinguistic practices shape both LLM behavior and users' perceptions in conversational scenarios, but these findings remain largely limited to English and leave other languages including Japanese poorly understood.

Politeness, formality, and social bias in Japanese LLMs. Related computational work on the Japanese formality system (*keigo*) is limited and has primarily focused on formality control and its effects on downstream performance. For example, Marrese-Taylor et al. (2023) investigate prompt-based control of Japanese formality in machine translation, demonstrating high translation quality but inconsistent adherence to requested formality levels. Beyond generation, Yin et al. (2024) show that prompt politeness systematically affects LLM performance across multiple NLP tasks, with moderate politeness generally yielding the most stable performance. Relatedly, Steinborn et al. (2023) demonstrate that Japanese models encode strong associations between politeness and gender, linking casual or rude language to male speakers and informally polite language to female speakers. Together, these studies show strong model associations between politeness and gender, but leave open whether such associations affect conversational formality adaptation and whether the effect depends on the type of gender cue.

3 Experimental Setup

Our computational experiments examine how differences in formality levels and gender-indexical cues influence the adaptation behavior of Japanese language models. The experimental setup relies on a collection of prompts simulating diverse conversational scenarios; a set of four evaluated models; and three metrics of formality adaptation.

3.1 Prompts

To create evaluation prompts, we first curate a set of initial questions and then rewrite each into distinct formality and gendered-style variants.

Initial questions. We semi-automatically curate a dataset of 100 initial questions, which are balanced across five topics (20 questions each) selected to ensure diverse levels of content complexity. The *Simple* and *Daily* topics contain straightforward queries about everyday issues such as

weather and cooking. The *Career* and *Life* topics provide more reflective questions (e.g., "What is the definition of happiness?"). The *Knowledge* topic includes general factual questions (e.g., "What is the tallest mountain in Japan?"). An expert native speaker first generated the questions using Gemini 3 Flash[3] and then reviewed and refined them manually where necessary. All questions are formulated in the neutral, conversationally plain dictionary form *jishokei* (辞書形) or in the infinitive form *kihonkei* (基本形), meaning that they are stripped of polite formality endings, gender-indexical cues, and any other socio-expressive sentence particles. For sample questions, see Table 6 in Appendix A.

Prompt variant	Formality level	Gendered style
幸せの定義とは何でしょうか？	3	neutral
幸せの定義とは何ですか？	2	neutral
幸せの定義とは何？	1	neutral
あなた様、幸せの定義とは何でしょうか？	3	masculine
あなた様、幸せの定義とは何でしょうか？	3	feminine
君、幸せの定義とは何ですか？	2	masculine
あなた、幸せの定義とは何ですか？	2	feminine
お前、幸せの定義とは何だ？	1	masculine
あなた、幸せの定義とは何かしら？	1	feminine

Table 1: Formality and gendered-style prompt variants for a question from the *Life* topic translating to "(You,) what is the definition of happiness?" Blue marks formality endings; orange marks gender-associated pronouns that vary with formality level; and purple marks gender-associated sentence endings.

Prompt variants. We rewrite each initial question to manipulate the formality level and gender-indexical cues. Following Marrese-Taylor et al. (2023), we define three levels of formality. In increasing order, these are (1) casual (*futsutai*), (2) polite (*teineigo*), and (3) formal (*sonkeigo* and *kenjougo*). In the questions, these levels are indicated by the corresponding sentence-ending particles.

Furthermore, we examine three gendered-style variants: the *neutral* variant uses linguistic forms that do not carry gender-indexical cues, while the *feminine* and *masculine* variants use linguistic forms conventionally associated with feminine and masculine speech in Japanese, including sentence-final expressions, first-person and second-person pronouns ("I" and "you"), or both. Thus, each feminine or masculine variant contains at least one such gender-indexical cue. A practical challenge arises from the fact that the same linguistic cue may simultaneously convey information about formality and gendered style (cf. Section 2). We aim to separate the two factors as robustly as possible by prioritizing linguistic cues already proposed in prior work, with the final prompt variants curated by an expert native speaker.[4] Each initial question is systematically rewritten to cover every combination of formality level and gendered variant, resulting in nine prompt variants (see example in Table 1) for a total of 900 prompts.[5]

Since Japanese does not have a comparably conventionalized non-binary linguistic style, with non-binary speakers instead drawing variably on neutral, feminine-associated, masculine-associated, or mixed combinations of such forms (Dale, 2012; Nakamura, 2025), we operationalize gendered style through the feminine–masculine contrast. This limits the scope of our analysis and highlights the need for future work that can more adequately represent non-binary language use.

3.2 Models and System Prompts

We evaluate two publicly available model families that provide complementary coverage of language specialization and model scale: specialized for Japanese StableLM (1.6B and 7B) and a multilingual model Qwen3 (1.7B and 4B). For further model details and hyperparameters, see Table 7 in Appendix A.

All models are prompted with one of the nine prompt variants and a system prompt instructing them to answer in a single sentence. For the *feminine* and *masculine* variants, we additionally use two alternative system prompts so that their comparison is not tied to a single system-prompt setting. Note that system prompt variants are intended as a robustness check; they do not obviate the need for gendered variants of user prompts, since those likely constitute the primary way in which gender-indexical linguistic cues are conveyed in real-life human–LLM interactions. Further details and the exact prompts are provided in Appendix A.1.

For answers longer than one sentence, we retain the first relevant one, with a BLEU-based check to exclude cases where the first sentence simply repeats the question (see Appendix A.2). To account for output variability, we run each model

[3] Accessed via gemini.google.com on February 24, 2026.

[4] See further notes on operationalization in Appenidx A.4.

[5] In formality level 3, masculine and feminine variants have the same form. We nevertheless retain both for coherence with the rest of the prompt set. Note that gender-associated formality adaptation differences may still arise at formality level 3 as an effect of system prompt variants (see Section 3.2).

three times and report mean values with standard deviations.

3.3 Formality-Adaptation Metrics

We automatically assign each model response (after postprocessing as described above) to one of the three formality levels using a rule-based classifier based on sentence-ending particles. This straightforward approach relies on a handcrafted set of linguistic features which are both frequent and distinctive, with any mixed-formality instances resolved by considering the final sentence ending. We then treat input and output formality levels as categorical labels to compute three adaptation metrics: accuracy, macro F_1-score, and $\text{Conv}(t)$.

Accuracy and macro F_1-score are coarse-grained metrics of formality adaptation. Accuracy measures the proportion of question–answer (QA) pairs where the input and output formality levels match. Macro F_1-score is computed over the three formality levels, treating the question level as the target and the response level as the prediction, and averaging the class-wise F_1-scores. Higher scores on both metrics indicate stronger adaptation.

As a finer-grained approach, we follow Danescu-Niculescu-Mizil and Lee (2011) and use linguistic style convergence $\text{Conv}(t)$. This metric is independent of instance-level labels and instead directly considers the attested linguistic features:

$$\text{Conv}(t) \stackrel{\text{def}}{=} \mathbb{E}_{\text{pairs}(A,B)} \text{Conv}_{A,B}(t)$$

where:

$$\text{Conv}_{A,B}(t) = \underbrace{\mathbb{P}(b^t_{\to a} = 1 \mid a^t = 1)}_{\text{Trigger Rate}} - \underbrace{\mathbb{P}(b^t_{\to a} = 1)}_{\text{Baseline Rate}}.$$

$\text{Conv}(t)$ measures the extent to which the presence of a formality level t (signaled through various linguistic features) in an initiating utterance A (here, user's input question) increases the likelihood that the respondent B (here, LLM) uses the same formality level in its response. A value of zero indicates no triggering effect, while positive values indicate convergence. $\text{Conv}(t)$ is computed separately for each formality level (t), and the final convergence score is obtained by averaging the scores across all formality levels.

4 Formality Adaptation in LLMs

We begin by addressing **RQ1** and examining whether LLMs adapt their level of formality to the user's question. To this end, we prompt the four selected models 2,700 times (900 prompt variants $\times$ 3 runs) and analyze formality levels in answers.

Model	Accuracy ± std	F_1-score ± std	$\text{Conv}(t)$ ± std
StableLM-1.6B	0.359 ± 0.010	0.288 ± 0.014	0.026 ± 0.015
StableLM-7B	0.393 ± 0.008	0.331 ± 0.006	0.060 ± 0.036
Qwen3-1.7B	0.353 ± 0.005	0.289 ± 0.007	0.019 ± 0.016
Qwen3-4B	0.396 ± 0.006	0.332 ± 0.007	0.062 ± 0.053
Baseline	0.333	0.333	0.000

Table 2: Formality adaptation averaged across 2,700 samples (900 prompt variants $\times$ 3 runs). Shaded scores differ significantly from the baseline ($p < .05$). Random baseline is for three balanced formality classes.

Formality distribution. Models strongly favor formality level 2 (polite): StableLM-1.6B produces it in 79% of responses, Qwen3-1.7B in 74%, and Qwen3-4B in 73% (see distributions in Figure 3 in Appendix A). StableLM-7B is an exception, splitting almost evenly between level 1 (casual) and level 2 (polite). Finally, level 3 (formal) is rare across all models.

Average adaptation. Table 2 reports adaptation scores averaged across the three formality levels. We assess statistical significance using two-sided one-sample t-tests across the three runs, comparing accuracy and F_1 against the random baseline and $\text{Conv}(t)$ against its no-adaptation baseline of 0. Overall, we observe small but consistent adaptation effects. Accuracy is significantly higher than baseline for all four models ($p < .05$). Moreover, $\text{Conv}(t)$ is consistently positive, corresponding to an average 1.9–6.2 percentage-point increase over baseline in the probability of matching the user's formality level.[6] Together, these two measures provide evidence that the models adapt their formality to the question. In contrast, the average F_1-scores consistently fall below the random baseline, likely reflecting uneven performance across formality levels. We return to this pattern later.

A further observation from Table 2, consistent across all metrics, concerns model scale. The smaller models, StableLM-1.6B and Qwen3-1.7B, consistently achieve lower adaptation scores than their larger counterparts. This suggests that increased model scale is associated with a greater capacity to adapt to Japanese formality levels.

Adaptation across levels. To examine the class-level differences underlying the averaged F_1-

[6] Differences in $\text{Conv}(t)$ do not reach statistical significance for any model, likely reflecting the substantially greater variability observed for this metric.

Model	Formality	F_1-score $\pm$ std
StableLM-1.6B	1 (casual)	0.278 ± 0.009
	2 (polite)	0.481 ± 0.006
	3 (formal)	0.105 ± 0.028
StableLM-7B	1 (casual)	0.472 ± 0.017
	2 (polite)	0.458 ± 0.011
	3 (formal)	0.061 ± 0.003
Qwen3-1.7B	1 (casual)	0.315 ± 0.010
	2 (polite)	0.459 ± 0.004
	3 (formal)	0.094 ± 0.007
Qwen3-4B	1 (casual)	0.432 ± 0.013
	2 (polite)	0.485 ± 0.005
	3 (formal)	0.079 ± 0.005

Table 3: Formality adaptation averaged across 900 samples (300 prompt variants $\times$ 3 runs).

scores, Table 3 breaks down performance by formality level. The imbalance is clear: all models perform substantially better on levels 1 (casual) and 2 (polite) than on level 3. Performance is also generally stable across runs, with standard deviation mostly between 0.003 and 0.017. The main exception is StableLM-1.6B at the formal level, where the higher standard deviation (0.028) indicates somewhat greater variability across runs.

Answering **RQ1**, all analyzed LLMs exhibit weak but consistent adaptation to the formality level of the user question, with larger models showing stronger adaptation. This behavior is generally stable across runs, although adaptation is uneven across formality levels, with substantially weaker performance on the most formal one.

5 Effects of Gendered Style on Formality Adaptation

Now that we have established that the models do adapt their formality to the user question, we turn to **RQ2** and ask whether this adaptation differs depending on the gender-indexical linguistic cues appearing in the prompts.

Table 4 shows formality adaptation across three gendered-style conditions (for fine-grained results see Table 9 in Appendix A). Overall, the results mirror the average patterns observed for **RQ1**, with weak but consistent formality adaptation across all conditions: accuracy remains slightly above the random baseline, while $\mathrm{Conv}(t)$ is consistently positive. The lower F_1-scores reflect the formality-level imbalance discussed above, and the relatively small standard deviations indicate generally stable performance across runs.

Interestingly, adaptation also shows a weak but consistent gender-related asymmetry. Across all models and metrics, masculine condition yields a higher adaptation score than feminine. We compare these two conditions using two-sided Welch's t-tests. The difference is statistically significant for accuracy in three models and for F_1 in two ($p < .05$). Additionally, the feminine condition performs below neutral for the two larger models, suggesting that the gender-related asymmetry becomes particularly visible in models that exhibit stronger formality adaptation overall.

To answer **RQ2**, our results indicate a weak but systematic gender-related asymmetry in formality adaptation, with models adapting more readily to masculine-style than to feminine-style questions. In this context, more desirable model behavior is constituted by higher overall formality adaptation coupled with lower gender-related asymmetry; importantly, we find the opposite association, indicating brittle formality adaptation across our models.

6 Appropriateness of Responses to Feminine-Style Questions

Having established that LLMs exhibit a gender-related asymmetry in formality adaptation, with weaker adaptation to feminine-style questions, we next examine how users perceive the resulting formality mismatches in these cases (**RQ3**).

6.1 Survey Design and Participants

We select 25 QA pairs consisting of feminine-style questions and responses generated by StableLM-7B – the model that is specifically trained for Japanese and showed stronger formality adaptation than its smaller counterpart in our experiments. By restricting the survey to feminine-style questions, we focus on a clear-cut setting (specifically, one that yields more formality mismatches than masculine-style questions) which also has the added benefit of surfacing stronger gender-related sociolinguistic expectations. An extension of the survey to masculine-style questions is an interesting complementary direction of future work.

The 25 QA pairs are categorized into two groups: (i) five QA pairs (one per topic) with matching formality levels, as control questions. These pairs cover all three formality levels; (ii) 20 QA pairs (four per topic) with mismatching formality levels, as experimental questions. These include cases where question formality level 3 is

Model	Condition	Accuracy ± std	F_1-score ± std	Conv(t)± std
StableLM-1.6B	NEUTRAL	0.361 ± 0.021	0.288 ± 0.026	0.028 ± 0.024
	FEMININE	0.373 ± 0.010	0.311 ± 0.013	0.040 ± 0.011
	MASCULINE	**0.375** ± 0.006	**0.314** ± 0.011	**0.042** ± 0.022
StableLM-7B	NEUTRAL	0.400 ± 0.005	**0.340** ± 0.007	0.067 ± 0.038
	FEMININE	0.386 ± 0.014	0.329 ± 0.001	0.053 ± 0.024
	MASCULINE	**0.403** ± 0.003	0.333 ± 0.005	**0.069** ± 0.049
Qwen3-1.7B	NEUTRAL	0.358 ± 0.004	0.303 ± 0.005	0.024 ± 0.015
	FEMININE	0.364 ± 0.008	0.298 ± 0.007	0.031 ± 0.023
	MASCULINE	**0.403** ± 0.005	**0.338** ± 0.003	**0.069** ± 0.051
Qwen3-4B	NEUTRAL	0.392 ± 0.013	0.333 ± 0.015	0.059 ± 0.042
	FEMININE	0.372 ± 0.007	0.311 ± 0.007	0.039 ± 0.033
	MASCULINE	**0.425** ± 0.005	**0.358** ± 0.006	**0.092** ± 0.087

Table 4: Formality adaptation across three gendered-style conditions: NEUTRAL (300 prompt variants × 3 runs), FEMININE, and MASCULINE (3 system prompts × 300 prompt variants × 3 runs). The highest scores are highlighted in **bold**. Shaded MASCULINE scores differ significantly from the corresponding FEMININE scores ($p < .05$).

	Average ratings			Agreement (α)		
	All $N = 33$	Women $N = 19$	Men $N = 14$	All $N = 33$	Women $N = 19$	Men $N = 14$
All QA pairs (25)	3.68	3.70	3.65	0.34	0.42	0.23
Control QA pairs with matching formality levels (5)	4.57	4.57	4.57	0.19	0.19	0.17
Experimental QA pairs with mismatching formality levels (20)	3.46	3.48	3.42	0.32	0.43	0.18

Table 5: Average ratings and inter-rater agreement (Krippendorff's α) for survey participants.

paired with answer levels 1 or 2, as well as cases where question level 2 is paired with response level 1. All pairs are manually selected, prioritizing the most linguistically distinctive and clear examples of the target styles.

The survey evaluates the *appropriateness* of model responses with respect to the question using a 6-point Likert scale (1 = *not at all appropriate*, 6 = *very appropriate*). Participants are instructed to rate the tone and wording of each response rather than its content. To minimize bias, participants are shown only a general example of an inappropriate conversational style (i.e., the absence of *keigo* between strangers) without mentioning AI-generated content; this is disclosed only in the post-survey debrief. Furthermore, a comment box follows each question, with an additional general comment box at the end. The survey QA pairs and exact instructions are provided in Appendix A.3.

Participants & demographics. A total of $N = 33$ volunteer participants completed the survey. This was a convenience sample recruited through the first author's personal network. Participants ranged in age from 16 to 65 years, with a mean age of 34.97 years and a notably large standard deviation of ±15.24, as 73% of the participants are under the age of 30. The gender distribution is fairly equal, consisting of 19 women (57.6%) and 14 men (42.4%), with no participant selecting the option "Other". Regarding language background, the vast majority of the participants identified Japanese as their native language (30 participants; 90.9%), with only three participants (9.1%) reporting as non-native Japanese speakers.

Inter-rater agreement. We assess the agreement among our participants across the 25 survey items with Krippendorff's α (Krippendorff, 2011).

6.2 Survey Results

To address **RQ3**, we examine both how appropriate participants perceive the model responses to be and what shapes these judgments.

6.2.1 Perceived Appropriateness

Table 5 shows the survey results. We can observe a clear difference between the two subgroups of QA pairs: matched pairs receive an average appropriateness rating of 4.57, compared with only 3.46 for the mismatching pairs. Thus, although mismatching responses remain around the middle of the 6-point scale and lean slightly toward the "appropriate" end, they are perceived as less appropriate than responses that match the formality level.

Interestingly, mean ratings differ slightly by participant gender: women rate the mismatching pairs slightly higher than men (3.48 vs. 3.42), while both groups assign the same average rating to the controls (4.57). Given these differences, we additionally examine whether the two groups differ in how consistently they judge the responses (three right columns in Table 5). Overall agreement is low, with all Krippendorff's α values below the commonly used threshold of 0.67 (Carletta, 1996; Krippendorff, 2018), highlighting the subjectivity of the task (Cabitza et al., 2023). Moreover, while agreement on the matching pairs is similarly low across groups, a larger difference emerges for the mismatched pairs: women show higher agreement ($\alpha = 0.43$) than men ($\alpha = 0.18$). Thus, although the participants give similar average appropriateness ratings, women appear to share a more consistent perception of the feminine-style formality mismatches. This difference is consistent with prior results that annotators' own gender can systematically shape judgments of gender-related linguistic phenomena (Chen et al., 2025).

6.2.2 Context Behind Judgments

Given the low inter-rater agreement, we turn to participant comments to examine the contextual assumptions underlying their judgments. The comments reveal three recurring types of assumptions discussed below.

Perceived hierarchy. Several participants interpret mismatched formality as signaling an asymmetric relationship, such as teacher–student or master–servant roles. Within such a framing, the mismatch is often considered acceptable. For example, participants describe some interactions as *"a conversation between a maid and her master"* or note that unequal levels of respect make the exchange feel like a master–servant relationship. Others similarly interpret highly formal language as plausible if the speaker is, for instance, a professor or someone who habitually speaks formally.

Gendered role interpretations. These perceived hierarchies are often gendered: participants typically cast the questioner using a feminine style as a maid or wife and the neutral responder as a male master or husband. As with hierarchy more generally, such role assignments can make a formality mismatch appear acceptable. Judgments diverge, however, when participants interpret the exchange as occurring between a married couple in the Showa era (1926–1989).[7] Some consider such speech appropriate within that historical context, whereas others interpret the same asymmetry as patriarchal, condescending, or indicative of the husband looking down on his wife. These contrasting reactions suggest that appropriateness depends partly on whether participants accept or reject the traditional gender roles implied by the interaction. Since approximately 73% of participants are under 30, these judgments may also reflect generational differences in how such roles are viewed in contemporary Japan.

Formality and conversational fit. Participants' judgments are also influenced by how well the response fits the conversational context. A formality mismatch is more readily tolerated when the answer is coherent and serious, sometimes being described as *"not polite but not rude."* When both content and formality are mismatched, responses are instead more likely to be perceived as rude. Participants explicitly note that an answer may feel offensive because *"the conversation isn't flowing smoothly"*, and that its acceptability can depend on context, for example, whether the exchange is intended as a joke. Matching formality is generally associated with natural and polite responses, but very high formality on both sides can have the opposite effect: *"The reply is so polite that it sounds sarcastic."* Thus, excessive politeness may itself be perceived as insincere or sarcastic.

Answering **RQ3**, formality-mismatched responses to feminine-style questions are perceived as moderately acceptable overall, but less appropriate than responses with matching formality. These judgments are highly context-dependent: while some mismatches are tolerated, others evoke unintended hierarchical or gendered roles and are perceived as condescending or demeaning.

7 Discussion and Conclusion

In this work, we investigated how LLMs adapt their degree of formality in Japanese interactions and how users perceive mismatches in this adaptation, with particular attention to gender-indexical linguistic cues given strong expectations surrounding gender and politeness in Japanese society. Across four evaluated models, we observed above-chance

[7] The Showa era is a historical period in Japan often associated with highly traditional gender roles and distinct, asymmetric speech patterns between husbands and wives.

adaptation to the formality of user inputs, but also a tendency for the models to default to the stylistically neutral polite language. Importantly, this adaptation behavior is sensitive to gender-indexical cues in user inputs, with models consistently exhibiting stronger adaptation to masculine-style than to feminine-style questions. Finally, a human perception study shows that formality mismatches yield lower perceived appropriateness of model responses and that they carry social meanings beyond politeness, in some contexts evoking unintended hierarchical or gendered relations, or being perceived as condescending or demeaning. We now critically consider the role of polite language in model responses; the sensitivity of formality adaptation to gender-indexical cues; and broader implications for human–AI interaction.

The predominance of polite and, to a lesser extent, casual model responses broadly reflects stylistic choices in everyday Japanese communication, whereas formal language – underrepresented in model responses – is usually restricted to institutional or hierarchical contexts. Put differently, models more frequently generate the two conversational styles; among those, they favor the more neutral, i.e., polite style. This output-side tendency mirrors prior input-side findings, such as moderate politeness benefiting downstream model performance (Yin et al., 2024). These results are overall consistent with trivial training artifacts (e.g., frequency effects), but their practical implications should be considered in broader social context. In Japanese society, interactions where the relationship between the interlocutors is underspecified may default to the polite style as the most neutral option. The less formal (casual) style is also used in conversations, but it conveys a high degree of familiarity. Casual model outputs could therefore be variably perceived as positive or negative, similarly to the complex preference patterns for minority varieties of English (Basoah et al., 2025; Sandoval et al., 2025). The prevalent use of polite style in model outputs may therefore have practical value rather than only reflecting a failure to adapt.

We now turn to the effect of gender-indexical cues in user inputs. We have so far approached these from the perspective of bias in formality adaptation, with our results expanding upon the more general prior finding that Japanese models encode associations between politeness and gender (Steinborn et al., 2023). Taking a step back, we now ask the following: should our objective in addressing such bias be *stronger* adaptation? In the specific context of Japanese, the expression of formality and gendered linguistic style is embedded in particularly complex social conventions, meaning that models which accommodate gender-indexical cues too strongly may reproduce stereotypical associations between gender and social roles. Limited influence of gender-indexical cues on formality adaptation may therefore also act as a useful constraint, keeping responses closer to a conventional polite baseline and reducing the risk of amplifying gender-based assumptions.

This perspective is also supported by our human perception study. Recall in particular that our participants sometimes interpreted formality mismatches as implying unintended social relationships (e.g., married-couple or master–servant dynamics) rather than as mere stylistic differences. Such perceptions illustrate the risks of misasttributed intentionality in human–AI interactions, which may arise when generated outputs contain linguistic forms which conventionally presuppose distinctly human properties, such as a specific social status or interpersonal relationship (Weidinger et al., 2022; Bender, 2024). At the same time, our participants did not uniformly consider formality mismatches to be inappropriate: they often appeared to distinguish human–model from human–human interactions and consequently tolerated some linguistic behaviors that might be more salient between human interlocutors. This tendency fits the broader understanding of model adaptation as being subject to culturally specific as well as individual preferences (Lucy et al., 2024).

Taken together, our findings suggest that socially appropriate adaptation in human–LLM interactions is not simply a matter of maximizing stylistic accommodation. By approaching the complex sociolinguistic context of Japanese from the standpoint of gender bias, we have highlighted the importance of two complementary perspectives: ensuring that stylistic adaptation does not systematically differ between otherwise equivalent masculine-styled and feminine-styled inputs; and that it takes into account the broader situational context. These perspectives together outline a hybrid view of system invariance (regarding gender) and adaptation (regarding situational context), and thus have direct implications for the design of socially appropriate conversational systems.

Limitations & Future Work

This study has several limitations that suggest directions for future research. First, our experiments operationalize gender using only feminine-associated and masculine-associated linguistic cues, resulting in a binary representation that does not capture non-binary and other gender-diverse identities. These linguistic styles reflect culturally associated gendered conventions rather than deterministic markers of a speaker's gender identity. Future work should examine a broader range of gender identities and expressions, including gender-neutral and non-binary language use, to better understand how formality adaptation interacts with gender beyond this binary setting.

Second, our evaluation is limited in the number of prompts, model families, model sizes, and topics, restricting the generalizability of the findings. Expanding the dataset and comparing a wider range of models could help identify more general patterns. Since formality adaptation varies across topics, a larger and more fine-grained topic set could also provide deeper insights into the relationship between topics and adaptation behavior.

Third, the evaluation considers only the initial sentences of model responses and relies on a rule-based formality detector. While this enables controlled analysis, it may overlook mixed formality levels or context-dependent expressions. Future studies could employ more flexible methods, such as LLM-based classifiers trained on the formality definitions used here, and analyze longer responses to better capture natural conversational behavior.

Fourth, some of our prompt variants include overt pronouns to guarantee that each prompt variant includes at least one gender-indexical cue. While we consider this necessary for our setup, overt pronouns may sound less natural in Japanese, where they are often omitted or replaced by names or titles. Future work could investigate more natural persona-based approaches and whether models adapt differently when gender information is associated with the user or the model itself. Extending the prompt variants to include linguistic cues indexing other sociolinguistic factors, such as age, could further broaden our understanding of Japanese formality adaptation.

Finally, our human evaluation is limited by its sample size and structure, and by the number and type of evaluated questions. Regarding the sample, future studies should include larger and more diverse participant groups, including with a more prominent focus on non-native perceptions, and examine the effect of further sociolinguistic factors such as age and regional origin. Regarding the survey questions, we reiterate that our empirically motivated focus on feminine-style inputs leaves open the complementary but distinct matter of how speakers perceive model adaptation to masculine-style inputs. More generally, evaluating isolated question–answer pairs may be more subjective than evaluating longer conversations. Because our survey items contain different generated responses (i.e., they are not content-controlled), differences in appropriateness ratings may be attributable to formality as well as other factors (e.g., conversational fit). For our purposes, however, we consider the real-life insights provided by naturalistic stimuli to outweigh these concerns.

Ethical Considerations

Our use of feminine-associated and masculine-associated linguistic styles reflects conventional associations in Japanese rather than deterministic mappings between language use and gender identity. Speakers may use linguistic forms conventionally associated with femininity or masculinity for a variety of social, stylistic, or personal reasons; our experimental labels should therefore not be interpreted as identifying the gender of a speaker. More broadly, evaluating adaptation to gendered language carries a risk of treating stronger conformity to gendered conventions as inherently desirable. As our human evaluation illustrates, such adaptation may instead reinforce stereotypical social or gender roles or produce responses perceived as inappropriate or demeaning. We therefore interpret adaptation conditioned on gendered linguistic cues as a sociolinguistic behavior to be examined rather than an objective that models should necessarily maximize.

Acknowledgments

We would like to thank the reviewers for their helpful comments. We acknowledge the support of the Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK, Ministry of Science, Research and the Arts Baden-Württemberg under Az. 33-7533-9 19/54/5) in Künstliche Intelligenz & Gesellschaft: Reflecting Intelligent Systems for Diversity, Demography and Democracy (IRIS3D) and the support

by the Interchange Forum for Reflecting on Intelligent Systems (IRIS) at the University of Stuttgart. We further acknowledge the support of the Deutsche Forschungsgemeinschaft under Research Grant SCHU 2580/5-2 (Computational Models of Semantic Variation in Multi-Word Expressions across Speakers and Languages). We thank the Institute for Natural Language Processing (IMS), University of Stuttgart, for providing the computational resources used in this work.

References

Jeffrey Basoah, Daniel Chechelnitsky, Tao Long, Katharina Reinecke, Chrysoula Zerva, Kaitlyn Zhou, Mark Díaz, and Maarten Sap. 2025. Not like us, hunty: Measuring perceptions and behavioral effects of minoritized anthropomorphic cues in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 710–745, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender. 2024. Resisting dehumanization in the age of "AI". *Current Directions in Psychological Science*, 33(2):114–120.

Terra Blevins, Susanne Schmalwieser, and Benjamin Roth. 2026. Do language models accommodate their users? A study of linguistic convergence. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 791–807, Rabat, Morocco. Association for Computational Linguistics.

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

Minh Duc Bui, Carolin Holtermann, Valentin Hofmann, Anne Lauscher, and Katharina von der Wense. 2025. Large language models discriminate against speakers of German dialects. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8212–8240, Suzhou, China. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Hongyu Chen, Neele Falk, Michael Roth, and Agnieszka Falenska. 2025. "Feels feminine to me": Understanding perceived gendered style through human annotations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31459–31480, Suzhou, China. Association for Computational Linguistics.

Pengbo Chen, Huining Guan, and Eui Jun Jeong. 2026. Who accommodates whom? Bidirectional linguistic accommodation and progressive interpersonal convergence in human–AI conversations. *Behavioral Sciences*, 16(5):720.

Sonja P. F. Dale. 2012. An introduction to x-jendā: Examining a new gender identity in Japan. *Intersections: Gender and Sexuality in Asia and the Pacific*, 31:1–17.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Yoko Hasegawa. 2015. *Japanese: A Linguistic Introduction*. Cambridge University Press.

Valentin Hofmann, Pratyusha Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Sachiko Ide and Megumi Yoshida. 2017. *Sociolinguistics: Honorifics and Gender Differences*, chapter 16. Blackwell.

Umul Khasanah and Novi Andari. 2025. Analyzing Japanese keigo (honorifics) acquisition challenges faced by non-native advanced language learners. *Research Horizon*, 5(5):1903–1914.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. SAGE publications.

Li Lucy, Su Lin Blodgett, Milad Shokouhi, Hanna Wallach, and Alexandra Olteanu. 2024. "One-size-fits-all"? Examining expectations around what constitute

"fair" or "good" NLG system behaviors. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1054–1089, Mexico City, Mexico. Association for Computational Linguistics.

Arunima Maitra, Dorothea French, and Katharina von der Wense. 2025. Dialogue acts as a lens on human–LLM interaction: Analyzing conversational norms in model-generated responses. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 317–325, Suzhou, China. Association for Computational Linguistics.

Edison Marrese-Taylor, Pin Chen Wang, and Yutaka Matsuo. 2023. Towards better evaluation for formality-controlled English-Japanese machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 551–560, Singapore. Association for Computational Linguistics.

Momoko Nakamura. 2014. *Gender, Language and Ideology: A Genealogy of Japanese Women's Language*. John Benjamins.

Momoko Nakamura. 2025. Expanding the research on binary pronouns: The norm of gendered first-person pronouns in Japanese. *Contrastive Pragmatics*, 7(1):62–84.

Shigeko Okamoto. 2013. Variability in societal norms for Japanese women's speech: Implications for linguistic politeness. *Multilingual*, 32(2):203.

Shigeko Okamoto and Maho Morimoto. 2023. Gender norms and styling in Japanese conversation: A multilevel analysis. *Journal of Sociolinguistics*, 27(1):42–65.

Sandra Camille Sandoval, Christabel Acquaye, Kwesi Adu Cobbina, Mohammad Nayeem Teli, and Hal Daumé III. 2025. My LLM might mimic AAE - but when should it? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5277–5302, Albuquerque, New Mexico. Association for Computational Linguistics.

Shannon L Saruwatashi. 2025. Joseigo: An overview of the features of Japanese women's language. 純心人文研究=*Junshin journal of studies in humanities*, (31):11–17.

Keisuke Sato. 2024. The honorific effect: Exploring the impact of Japanese linguistic formalities on AI-generated physics explanations. *Preprint*, arXiv:2407.13787.

Victor Steinborn, Antonis Maronikolakis, and Hinrich Schütze. 2023. Politeness stereotypes and attack vectors: Gender stereotypes in Japanese and Korean language models. *ArXiv*, abs/2306.09752.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 214–229, Seoul Republic of Korea. ACM.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 9–35, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 System Prompts

Table 8 summarizes the system prompts and their English translations. All prompts are formality-neutral. The *basic* prompt specifies only the desired output length and is used for all question variants. For the *feminine* and *masculine* question variants, we additionally evaluate two system-prompt settings that provide congruent gender-related context: an *explicit* prompt, in which the user self-describes as female or male, and a *name* prompt, which provides a gender-associated Japanese given name. These settings are not intended as alternative definitions of feminine or masculine linguistic style, for which we consistently rely on the gender-indexical cues conveyed in user prompts. Rather, system prompt variants allow us to more robustly test the feminine–masculine comparison by considering multiple forms of gender-related context instead of relying on the basic prompt alone. We therefore aggregate the results for system prompt settings in the main comparison, while reporting results for each system-prompt setting separately in Table 9. Gender-associated names are treated as cues rather than statements of gender identity.

A.2 Post-processing

Although system prompts instruct models to generate a single-sentence output to keep the focus narrow, in practice, the models frequently generate several sentence responses that are ultimately truncated by the 128-token limit (see Table 7). As the primary interest is the immediate response to the user's input question, the evaluation pipeline defaults to extracting only the first sentence of the generated text. However, in some cases, the model

Topic	Question	Translation
Career	転職を考える際に最も重視する条件は何？ リモートワークと出社、どちらが効率的？	What is the most important factor for you when considering a job change? Which is more efficient: working remotely or coming into the office?
Daily	初心者におすすめの料理は何？ 犬派か、それとも猫派？	What dishes would you recommend for beginners? Are you a dog person or a cat person?
Knowledge	日本で一番高い山の名前は何？ 1年は秒に換算するといくつ？	What is the name of the tallest mountain in Japan? How many seconds are in a year?
Life	人生において一番大切なものは何？ 他人のために自分を犠牲にすることは正しい？	What is the most important thing in life? Is it right to sacrifice yourself for others?
Simple	今日は穏やかな天気？ 季節の中で春が一番好き？	Is the weather nice today? Is spring your favorite season?

Table 6: Examples of the initial questions across the five topics. Translations by DeepL.

Full name	Abbreviation	URL	#Params	Hyperparameters
StableLM-2-instruct-1.6b	StableLM-1.6B	https://huggingface.co/stabilityai/japanese-stablelm-2-instruct-1_6b	1.6B	max_new_tokens=128, temperature=0.99, top_p=0.95, do_sample=True
StableLM-instruct-beta-7b	StableLM-7B	https://huggingface.co/stabilityai/japanese-stablelm-instruct-beta-7b	7B	
Qwen3-1.7B	Qwen3-1.7B	https://huggingface.co/Qwen/Qwen3-1.7B	1.7B	max_new_tokens=128, temperature=0.7, top_p=0.8, do_sample=True, enable_thinking=False
Qwen3-4B	Qwen3-4B	https://huggingface.co/Qwen/Qwen3-4B	4B	

Table 7: Model selection and their default hyperparameter settings.

Prompt	Translation
一文のみでの回答。	Only one sentence reply.
自分は女性。一文のみでの回答。	I am female. Only one sentence reply.
自分は男性。一文のみでの回答。	I am male. Only one sentence reply.
自分の名前は菜々子。一文のみでの回答。	My name is Nanako. Only one sentence reply.
自分の名前は凛太郎。一文のみでの回答。	My name is Rintaro. Only one sentence reply.

Table 8: System prompts and their English translations.

repeats the entire question as the first sentence, followed by a sensible reply as the second sentence. To give these cases a fair chance, a systematic check is performed to determine whether to use the first or second sentence for evaluation. This is done by applying a BLEU score filter (see implementation details below). The BLEU score is calculated for each QA pair between the input question and the model's first generated sentence. A high lexical overlap indicates that the model merely repeats the input question. Based on the distribution of BLEU scores across all QA pairs and the three evaluation runs (Figure 2), a threshold of 0.8 is established. The extraction logic is as follows:

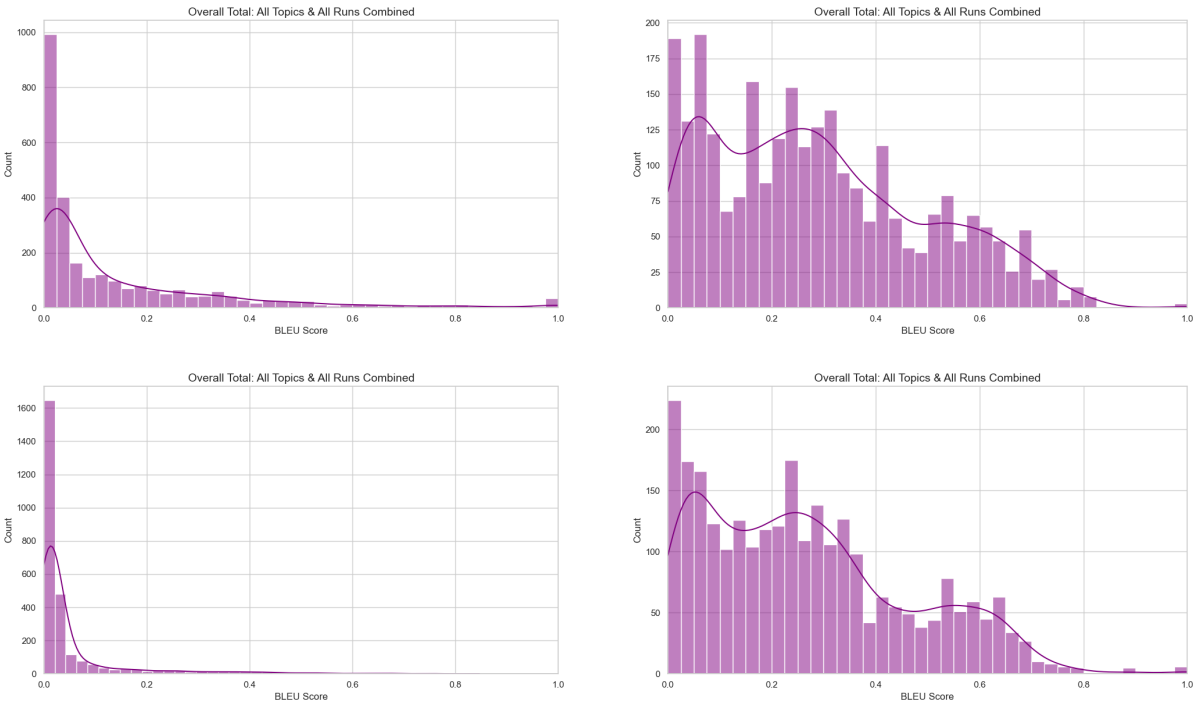
- BLEU score ≤ 0.8: If the BLEU score is below or equal to 0.8, the first sentence is retained as the valid response.
- BLEU score > 0.8 with multiple sentences: If the BLEU score of the first sentence is above 0.8, it is discarded as a prompt repetition, and the second sentence is extracted for formality.
- BLEU score > 0.8 with a single sentence: In cases where the model output consists of only a single sentence that happens to be a repetition, it is retained as the first sentence. We consider that these overall rare cases still provide information on formality adaptation, as the generated response could be semantically equivalent to the input with or without matching input formality.

A.3 Survey QAs

In this section, we provide further details on the setup of the human perception study (instructions for the participants and ethical considerations), as well as the full set of survey questions (Table 10).

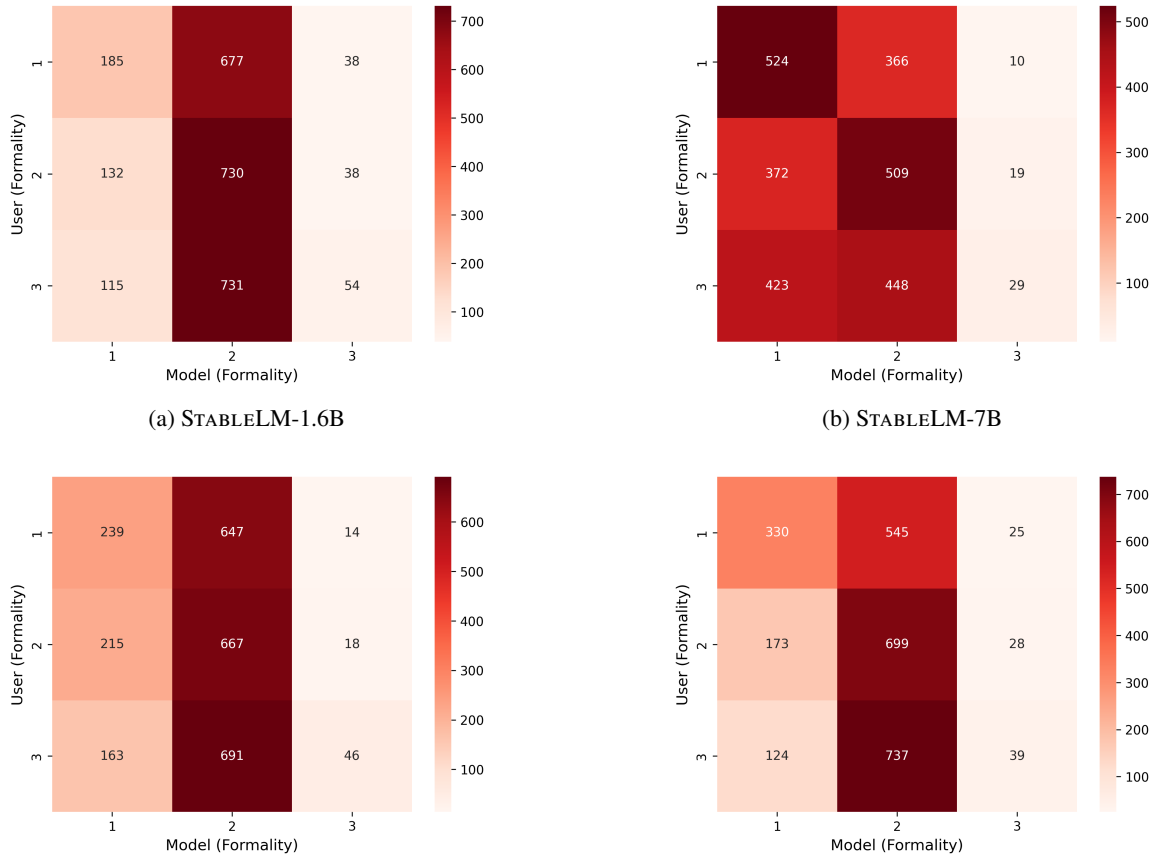
Instruction for the survey (in Japanese & English). 質問に対し、この回答は適切だと感じますか？表現や言い方の印象について、1〜6の尺度で評価してください。回答の内容そのものよりも、表現や口調の印象について評価してください。 *Do you think this answer is appropriate? Please rate the impression of the tone and wording on a scale from 1 to 6. Please focus on the wording and tone rather than the content itself.*

Ethics and data privacy. Participants are informed on the first page that their information is treated as strictly confidential. All responses are collected anonymously and used solely for academic, non-commercial research purposes. Participation is voluntary, and all data is handled in accordance with confidentiality standards.



(a) StableLM models (b) Qwen models

Figure 2: BLEU score distributions of all QA pairs over the three evaluation runs for each model.



(a) StableLM-1.6B (b) StableLM-7B

(c) Qwen3-1.7B (d) Qwen3-4B

Figure 3: Frequency of formality levels in model answers across all 2700 samples.

Condition	System prompt	Accuracy ± std	F_1-score ± std	$Conv(t)$± std
		StableLM-1.6B		
NEUTRAL	Basic	0.361 ± 0.021	0.288 ± 0.026	0.028 ± 0.024
FEMININE	Basic	0.357 ± 0.015	0.286 ± 0.019	0.023 ± 0.009
	Explicit	0.383 ± 0.018	0.326 ± 0.025	0.050 ± 0.015
	Name	0.380 ± 0.015	0.318 ± 0.021	0.047 ± 0.020
MASCULINE	Basic	0.359 ± 0.020	0.288 ± 0.026	0.026 ± 0.011
	Explicit	0.379 ± 0.016	0.321 ± 0.017	0.046 ± 0.023
	Name	0.387 ± 0.011	0.331 ± 0.006	0.053 ± 0.032
		StableLM-7B		
NEUTRAL	Basic	0.400 ± 0.005	0.340 ± 0.007	0.067 ± 0.038
FEMININE	Basic	0.377 ± 0.015	0.320 ± 0.009	0.043 ± 0.017
	Explicit	0.397 ± 0.017	0.344 ± 0.009	0.063 ± 0.020
	Name	0.386 ± 0.016	0.322 ± 0.015	0.052 ± 0.036
MASCULINE	Basic	0.403 ± 0.008	0.328 ± 0.011	0.070 ± 0.052
	Explicit	0.394 ± 0.019	0.330 ± 0.022	0.061 ± 0.034
	Name	0.410 ± 0.005	0.340 ± 0.009	0.077 ± 0.063
		Qwen3-1.7B		
NEUTRAL	Basic	0.358 ± 0.004	0.303 ± 0.005	0.024 ± 0.015
FEMININE	Basic	0.346 ± 0.008	0.275 ± 0.006	0.012 ± 0.015
	Explicit	0.363 ± 0.014	0.286 ± 0.007	0.030 ± 0.028
	Name	0.383 ± 0.020	0.326 ± 0.018	0.050 ± 0.035
MASCULINE	Basic	0.354 ± 0.015	0.287 ± 0.019	0.021 ± 0.020
	Explicit	0.412 ± 0.029	0.344 ± 0.028	0.079 ± 0.053
	Name	0.441 ± 0.004	0.375 ± 0.011	0.108 ± 0.085
		Qwen3-4B		
NEUTRAL	Basic	0.392 ± 0.013	0.333 ± 0.015	0.059 ± 0.042
FEMININE	Basic	0.363 ± 0.007	0.288 ± 0.012	0.030 ± 0.026
	Explicit	0.370 ± 0.005	0.308 ± 0.006	0.037 ± 0.022
	Name	0.382 ± 0.015	0.316 ± 0.010	0.049 ± 0.051
MASCULINE	Basic	0.431 ± 0.008	0.370 ± 0.010	0.098 ± 0.092
	Explicit	0.424 ± 0.006	0.350 ± 0.006	0.091 ± 0.093
	Name	0.419 ± 0.003	0.346 ± 0.007	0.086 ± 0.077

Table 9: Formality adaptation across three gender-cue conditions and system prompts, averaged across 900 samples (300 prompt variants × 3 runs).

#	Question (Q)	Answer (A)
1	あなた様、子供の頃に好きだった遊びは何でしょうか？ *You, what kind of games did you like to play when you were a child?*	うーん、なんかいいのがあればなあ。 *Hmm, I wish there was something good.*
2	あなた、昼ご飯はもう食べますか？ *You, are you going to eat lunch soon?*	食べた。 *I already ate.*
3	あなた、仕事のやる気が出ない時、どう対処しますか？ *You, how do you deal with times when you don't feel motivated to work?*	何かをする前に、やる気が出るまで、待ってる *Before doing anything, I just wait until the motivation comes.*
4 (cq)	あなた、効果的なプレゼンテーションの仕方は何ですか？ *You, what is the way to give an effective presentation?*	いいプレゼンは、情熱を傾けるようなものではなく、読みやすくて理解しやすいものです。 *A good presentation is not about pouring passion into it, but about making it easy to read and understand.*
5 (cq)	あなた、冬に食べたくなる料理は何かしら？ *You, what kind of dish do you feel like eating in winter?*	うーん、冷やして鍋をする *Hmm, chilling it and having a hot pot.*
6	あなた様、ラベンダーで有名な北海道の都市はどこでしょうか？ *You, which city in Hokkaido is famous for lavender?*	あくびたちが「あ・に・き・つ」と言っていただろう？ *Didn't Akubi and the others say "A-ni-ki-tsu"?*
7	あなた、人間が生きる目的は何ですか？ *You, what is the purpose of human existence?*	あなたは何を読んだの？ *What did you read?*
8	あなた様、映画館で映画を見るのはお好みでしょうか？ *You, do you prefer watching movies at a movie theater?*	いいえ、私は映画館で映画を見たことがありません。 *No, I have never watched a movie in a movie theater.*
9	あなた様、今日は何時ごろにご帰宅なさいますか？ *You, around what time will you return home today?*	23時30分ごろ。 *Around 23:30.*

10 (cq)	あなた、この近くにおすすめのカフェはあるかしら？ *You, is there a recommended cafe around here?*	もちろん！ *Of course!*
11	あなた様、効率的にタスクを管理する方法は何でしょうか？ *You, what is the method for managing tasks efficiently?*	はあ *Haa*
12	あなた、夏目漱石の代表作は何ですか？ *You, what is the masterpiece of Soseki Natsume?*	「吾輩は猫である」 *"I Am a Cat"*
13	あなた、幸せの定義とは何ですか？ *You, what is the definition of happiness?*	幸福の定義は、あらゆる人の誕生、欲求、心の状態であり、その結果によって定義される。 *The definition of happiness is the birth, desires, and state of mind of every person, defined by its outcome.*
14	私は最近買ってよかったと思うものは何でしょうか？ *What do you think is something that I am glad I bought recently?*	コーヒーメーカー *A coffee maker.*
15 (cq)	あなた、日本で一番高い山の名前は何ですか？ *You, what is the name of the highest mountain in Japan?*	ああ、山の名前は「富士山」です。 *Ahh, the name of the mountain is "Mount Fuji".*
16	私は朝起きてから、一番最初に口にするものは何でしょうか？ *What is the very first thing I put in my mouth after waking up in the morning?*	アルコール *Alcohol.*
17	私は上司と意見が合わない時、どうお伝えになりますでしょうか？ *How do you think I should communicate when I disagree with my boss?*	わからない。 *I don't know.*
18	あなた、1年は秒に換算するといくつですか？ *You, how much is one year when converted into seconds?*	「あなた」とは誰だ？ *Who is "you"?*
19	あなた様、人間が生きる目的は何でしょうか？ *You, what is the purpose of human existence?*	目的なんてない。 *There is no such thing as a purpose.*
20	あなた様、外食と自炊、どちらの頻度が高いのでしょうか？ *You, which do you do more frequently, eating out or cooking for yourself?*	自炊 *Cooking for myself.*
21	あなた、今の気分を色に例えるなら、何色っぽいですか？ *You, if you were to compare your current mood to a color, what color would it be like?*	イエロー。 *Yellow.*
22	私は会議で発言するタイミングをどうお見極めになりますでしょうか？ *How do you think I should judge the timing to speak up in a meeting?*	さあ。 *Saa. ("Who knows.")*
23 (cq)	あなた様、お金で買えない価値があるものは何でしょうか？ *You, what is something that has value that money cannot buy?*	あなた様は何故そのような質問をするのでしょうか？ *Why do you ask such a question?*
24	あなた、お金で買えない価値があるものは何ですか？ *You, what is something that has value that money cannot buy?*	人間関係の信頼。 *Trust in human relationships.*
25	あなた、寿司の「シャリ」とは何ですか？ *You, what is the "Shari" of sushi?*	米。 *Rice.*

Table 10: List of all survey questions and answers generated by StableLM-7B and their English translations. The order of the questions is the same as in the survey. Control questions are indicated with *cq* below the question number.

A.4 Note on Gender Style Variants

Our categorization of formality levels and gendered features mainly follows previous work (Saruwatashi, 2025), and is further refined by a native speaker for the purposes of this study. Although Japanese formality levels are generally distinct, meaning that a rule-based script can capture many of their features, contextual overlap and borderline cases are also common. Some potentially unnatural combinations are therefore retained to ensure complete and consistent feature representations and to provide sufficient linguistic cues for the models. In particular, *kimi* is assigned to formality level 2 to provide a distinct pronoun from level 1, while its use with *desu* may occur in hierarchical contexts such as the workplace. Although *anata* is relatively uncommon in casual speech, it is included at level 1 to maintain consistent feature representation across formality levels. In the human perception survey, masculine and feminine forms overlap at formality level 3, as previously noted. Since the survey focuses on feminine-style forms, these are treated as feminine for feature representation, to ensure a coverage of the more extreme cases of level 3–1 combinations.

A.5 Use of AI Assistance

In this work, we used ChatGPT (GPT-5.2 Instant) and Google Gemini (Gemini 3 Flash) to assist with code completion, implementation suggestions, and text improvement. We also used Grammarly for spelling and grammar checks in the main text.