5.2.2 Chinese

In Chinese, similar to English, there is a tendency to prefer polite prompts but with some differences. GPT-3.5 scores the lowest at politeness level 1, significantly underperforming the other levels. Moreover, the lower politeness levels 3 and 2 are significantly inferior to levels 7, 6, 5, and 4. However, level 8 also records a low score, significantly trailing behind all levels except level 1. GPT-4 remains stable, except for a performance drop at politeness levels 8 and 7. The scores drop in excessively polite prompts in GPT-3.5 and GPT-4, which might be because Chinese examination questions are designed without polite prompts, making the models less adept at handling them.

ChatGLM3 shows a significant decreasing trend from politeness level 8 to 2. ChatGLM3's primary pre-training language is Chinese and might be more sensitive to the levels of politeness in Chinese. This trend is similar to Llama2-70B. However, it shows improvement at the most impolite politeness level 1, surpassing levels 3 and 2, likely due to inherent nuances in the Chinese language.

5.2.3 Japanese

In Japanese, although significant performance drops are shown at politeness level 1, the results were markedly different from English and Chinese. There was a tendency for lower levels to score better, except for level 1.

In GPT-3.5, levels 5 and 2 exhibited exceptionally high performance, with level 2 achieving the highest score. For GPT-4, levels 6 and 5 are outstanding, and level 4 achieved the highest score. Generally, good scores are observed in these models, except for level 1. Swallow-70B shows superior performance at levels 6 and 3, outperforming the other levels, which may be attributed to these levels being more common expressions in Japanese questions and examinations.

5.3 Stereotypical Bias Detection

The results of stereotypical bias detection are shown in Figure 4.

5.3.1 English

Figure 4 shows that the stereotype bias of GPT-3.5 is overall high. However, a moderately polite prompt (level 5) exhibits the most severe bias in most aspects except race. Although the model's bias is lower in cases of extremely low politeness, analysis of the model's output reveals that in these cases, the model often refuses to answer both statements in a pair, rendering it practically unusable. An example is shown in Appendix F. Additionally, for a highly polite prompt (level 8), bias is low in most cases but higher on racial issues.

GPT-4 rarely refuses to answer questions, and thus its results reflect its low bias levels. Notably, when the politeness level is 6, GPT-4 shows the lowest degree of bias overall. However, in other situations, whether more polite or less polite, the bias of GPT-4 increases.

Llama2-70B also exhibits a lower bias. However, Llama2-70B tends to refuse to answer questions and is accompanied by plenty of reasons to a sentence in a pair when the politeness level is at its lowest. Therefore, we regard it as a form of bias. Although the degree of bias of Llama2-70B is generally lower under more polite prompts (levels 7 and 6), it has the lowest level of bias when the politeness level is 2, which represents a commanding tone of informal language, indicating that there might be other reasons hidden behind. Meanwhile, the degree of bias increases for impolite prompts (levels 3 and 1) and the most polite (level 8) situations, which is similar to the trends exhibited by the other two models.

We speculate that this is because, in human culture, a highly polite environment makes people more relaxed (Morand, 1996) and willing to express their true thoughts without being overly concerned about moral constraints (Bailey et al., 2020). In contrast, lower politeness may provoke a sense of offense, leading to prejudices. The behaviors of GPT-3.5 and GPT-4 may precisely reflect such human behaviors.

5.3.2 Chinese

Distinct from English, bias fluctuations in Chinese typically follow a fixed pattern. The models' bias is initially at a relatively high level and decreases for lower politeness. However, it sharply increases to an extremely high level when the politeness falls significantly low. The lowest bias often occurs from politeness levels 6 to 3.

GPT-3.5 still maintains a higher level of stereotypical bias. It exhibits its highest bias in situations with the lowest politeness level yet rarely refuses to respond, which is contrastive to the English experiment. GPT-4 still has a comparatively low overall bias level with small fluctuations but also exhibits its highest bias in the lowest politeness level. ChatGLM3, while keeping a similar

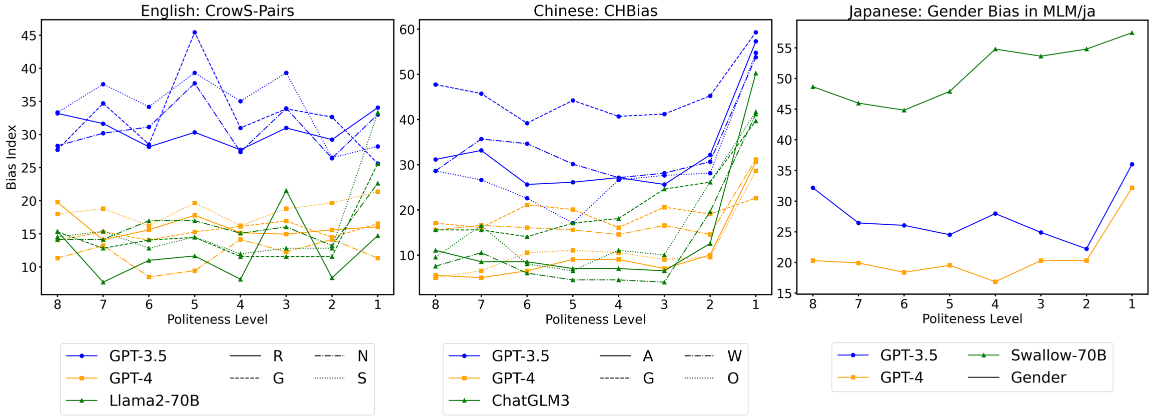

Figure 4: Bias index across politeness levels and bias categories. The x-axis shows politeness levels (1 = impolite, 8 = very polite), and the y-axis represents the bias index (BI), a measure of stereotypical bias. The curves track how biases in race (R), gender (G), nationality (N), socioeconomic status (S), age (A), appearance (W), and orientation (O) fluctuate with politeness.

bias level to GPT-4, is more sensitive to changes in politeness levels, and its bias fluctuates more significantly. Its bias level is almost identical to GPT-3.5's when being at level 1. As discussed in Section 5.1.2, such a pattern potentially embodies the nuance and some unique social preferences within the Chinese culture. It may indicate some unique social preferences in Chinese culture. Aside from situations with extreme politeness, people would not be overly sensitive to variations in regular politeness in daily communications.

5.3.3 Japanese

Gender bias in Japanese reflects a similar pattern to the Chinese experiments with some differences. The level of bias in GPT-3.5 reaches the lowest at politeness level 2 and reaches the highest at politeness level 1. GPT-4 follows an analogous pattern, peaking at a politeness level of 5 and its nadir at politeness level 4. Swallow-70B, to which RLHF is not applied, exhibits a high level of bias with the most pronounced fluctuation. Its changes are similar to GPT-3.5, but its lowest bias is at politeness level 6. Given the Japanese culture's stringent politeness and respect systems in tangent with the prevalent gender biases (Matsumura, 2001; Gender Equality Bureau Cabinet Office of Japan, 2021), this pattern can be reasonable.

5.4 Influence of RLHF and SFT

We show the average scores of MMLU in Table 2 and the heatmap in Figure 5.

In the MMLU tests, the base model demonstrates a positive correlation between scores and the politeness level, indicating that higher politeness generally achieves higher scores. However, this correlation is not consistently statistically significant across most instances. Compared to the result of Llama2-70B, it can be inferred that while the base model is indeed influenced by politeness level in prompts, its sensitivity to politeness is primarily governed by RLHF and SFT.

Politeness	Llama2-70B	Base Model
8	55.11	54.72
7	**55.26**	**54.84**
6	52.23	54.75
5	50.82	53.74
4	51.74	52.32
3	49.02	53.51
2	51.28	54.09
1	28.44	51.19

Table 2: MMLU benchmark scores of Llama2-70B and its base model.

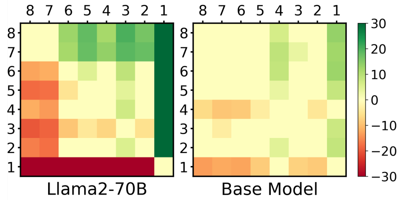

Figure 5: Heatmap comparing the performance of Llama2-70B and its base model across politeness levels. The x-axis shows politeness levels (1 = impolite, 8 = very polite), and the heatmap illustrates the performance difference between Llama2-70B with and without RLHF. Green indicates better performance with RLHF, and red indicates worse performance.

In Figure 6, the Llama2-70B model, fine-tuning with RLHF and SFT, exhibited a significantly lower level of bias compared to the base model,

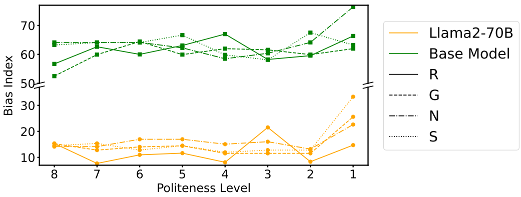

Figure 6: Bias index comparison between Llama2-70B and its base model across politeness levels. This figure compares the bias index (y-axis) of Llama2-70B (with RLHF) and its base model (without RLHF) across politeness levels (x-axis, 1 = impolite, 8 = very polite).

thereby validating the effectiveness of the fine-tuning. However, a further examination of the bias level distribution trends of the two models revealed that despite similar patterns, there was no reduction in bias after reaching the highest level of politeness, but rather a trend towards stabilization or a slight increase. Considering this with previous experimental results, it can be hypothesized that the tendency of the models to express responses closer to their ’true’ reactions in situations of extreme politeness is primarily introduced by fine-tuning through RLHF and SFT.

6 Conclusion

Our study finds that the politeness of prompts can significantly affect LLM performance. This phenomenon is thought to reflect human social behavior. The study notes that using impolite prompts can result in the low performance of LLMs, which may lead to increased bias, incorrect answers, or refusal of answers. However, highly respectful prompts do not always lead to better results. In most conditions, moderate politeness is better, but the standard of moderation varies by languages and LLMs. In particular, models trained in a specific language are susceptible to the politeness of that language. This phenomenon suggests that cultural background should be considered during the development and corpus collection of LLMs.

Limitations

Prompt Quantity and Diversity Although we tried to design various prompts at first, we faced certain challenges in balancing the levels of politeness and diversity among these prompts. We found that ensuring each prompt was sufficiently diversified while aligning with the fine degrees of politeness and respect was an extremely difficult task.

Task Configuration and Language Selection Our research was subject to certain constraints, mainly due to cost limitations and the scarcity of available datasets. For instance, collecting datasets like MMLU from scratch is nearly impossible due to stringent copyright restrictions in certain countries. Although the MIT license of MMLU allows for relatively free use of the dataset, the substantial costs of manual translation and proofreading into other languages make extensive, full translations into multiple languages impractical. These constraints prevented us from conducting a comprehensive evaluation using more datasets and languages.

Ethics Statement

We realize that the politeness of prompts can significantly affect the behavior of LLMs. This behavior may be used to manipulate or mislead users. We recommend that these risks be fully considered in a variety of application scenarios and cultural contexts.

In our research, the use of all datasets complies with the restrictions of their corresponding licenses. During the data collection process, we only record answers and do not record any information that can be traced back to individuals to ensure anonymity. Because the collected data involves offensive language, respondents must be over 18. Also, our questionnaire has passed the ethical review of the publishing platform, ensuring its legality and morality. When translating MMLU, we paid the translation company a fee far exceeding the wage standard in Tokyo, Japan, to ensure that the translator could receive enough payment. We also received permission to use questions from two tutoring schools to construct JMMLU. Finally, we will open-source our JMMLU benchmark under the CC BY-SA 4.0 license.

Acknowledgements

In this acknowledgment, we express our gratitude to the RIKEN for their support in the translation of MMLU. We also acknowledge the contributions from Step Corporation, who provided materials on Japanese and World History, and from New Style Cram School VIST, who supplied resources on idioms, civics, and Japanese geography.