Are large language models affected by politeness? Focusing on request speech acts in Korean

Gayeon Jung, Joeun Kang, Fei Li, Hansaem Kim

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

© 2024. Gayeon Jung, Joeun Kang, Fei Li, Hansaem Kim. Are large language models affected by politeness? Focusing on request speech acts in Korean. In Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.), *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation* (PACLIC 38), 894-903. Institute for the Study of Language and Information, Kyung Hee University. This work is licensed under the Creative Commons Attribution 4.0 International License.

Are large language models affected by politeness? Focusing on request speech acts in Korean

Gayeon Jung, Joeun Kang, Fei Li, Hansaem Kim
Interdisciplinary Graduate Program of Linguistics and Informatics
Yonsei University
Seoul, South Korea
`{wjdrkdus98, j0eun, feili0820, khss}@yonsei.ac.kr`

Abstract

This study examined the influence of politeness on large language models (LLMs) based on request speech acts in Korean, which features a highly developed system of polite expressions. To address this issue, we designed five levels of request prompts ranging from informal to highly formal on the basis of the politeness expression system of the Korean language. We then analyzed the responses of GPT-4, CLOVA X, Mixtral, and Solar to these prompts in terms of accuracy and friendliness. Relatively larger models, such as GPT-4 and CLOVA X, were sensitive to the politeness levels of the prompts. Furthermore, CLOVA X demonstrated an increase in accuracy and friendliness with the increase in the level of politeness of the prompts. In contrast, relatively smaller models, such as Mixtral and Solar, did not exhibit a consistent correlation between politeness and response quality. These findings indicate that the quantity of training data and the scale of the model are significant factors in discerning the nuances of language. They also highlighted the importance of considering politeness when designing Korean-specific prompts. Additionally, this study underscores the need to conduct an in-depth examination of the ability of LLMs to recognize politeness in diverse linguistic and cultural contexts.

1 Introduction

Recent advancements in artificial intelligence (AI) and natural language processing (NLP) have led to a surge in interest in human–computer interactions. Consequently, many studies have proposed that AI behavior should be designed to emulate that of humans (Priya et al., 2024, Lykov et al., 2024, Almeida et al., 2024). Linguists have posited that politeness represents a fundamental aspect of human language, which is pivotal in establishing social order (Li et al., 2023; Brown, 1987). Humans are susceptible to politeness during communication (Yin et al., 2024; Dillon, 2003). For example, human beings generally tend to assist others when requested in a polite language, but they tend not to cooperate when the request is made via an impolite language. In other words, the acceptance of a request is typically determined by the degree of politeness. These results demonstrate that politeness substantially impacts the capacity of the speaker to attain their objectives.

Korean is one of the few languages with an elaborate and explicit honorific system known as 경어법 (*gyeongeobeop*) (Lee, 1982). In Korean, the appropriate level of honorifics is systematically realized at multiple levels for all persons in a conversation, which results in an honorific system that differs from those of other languages (Han, 1999).

The current study examines the influence of the degree of politeness on large language models (LLMs) in request speech acts in Korean on the basis of the argument that AI behavior should mimic human behavior. Thus, it poses the following research questions:

- RQ1. Does the politeness of a prompt influence the response of LLMs?
- RQ2. If RQ1 is true, then how do LLMs differ in perceived politeness?
- RQ3. Why should politeness (not) be considered when designing prompts?

2 Related Work

2.1 Prompting

Prompts are inputs to the generative AI that guides the outputs of a model (Schulhoff et al., 2024; Meskó, 2023; White et al., 2023; Heston and Khun, 2023; Hadi et al., 2023; Brown et al., 2020). The advent of generative AI has motivated several studies to investigate effective prompting techniques to enhance the quality of model responses.

Schulhoff et al. (2024) established a systematic understanding of prompts by categorizing prompting techniques and analyzing their applications. The authors intended to provide a comprehensive understanding of prompts by discussing more than 200 prompting techniques, constructing a framework on them, and considering safety and security issues when utilizing them. This research is significant, because it provides a well-structured organization of the prompting techniques developed to date. Alternatively, Bsharat et al. (2023) introduced 26 fundamental principles for the organization of prompts to facilitate the efficient interaction of developers and general users with LLMs. The study evaluated the effectiveness of these principles on seven LLMs and demonstrated that the efficient reconstruction of prompt contexts improves the relevance and objectivity of responses. Notably, however, the methodology has been verified only for English. By providing an overview of prompts, Sahoo et al. (2024) addressed the lack of systematic organization and comprehension of prompt engineering methodologies. The study summarized the methods associated with 29 prompt engineering techniques, which offered insights into the advantages and disadvantages of each method.

A number of studies have explicitly focused on politeness in prompts. For example, Yin et al. (2024) evaluated the impact of politeness levels on LLMs in English, Chinese, and Japanese. The researchers observed that using impolite prompts typically results in suboptimal performance; nevertheless, excessively polite language does not ensure superior outcomes. Thus, the authors argued that politeness levels that yield the best performance vary across languages. This result demonstrated that LLMs mirror human behavior and are influenced by linguistic nuances in diverse cultural contexts. In a related study, Vinay et al. (2024) conducted an experiment on misinformation generated by LLMs using prompts that feature politeness and impoliteness. The finding illustrated that LLMs generate misinformation on the basis of subtle emotional understanding in polite prompts. Conversely, with impolite prompts, LLMs refrain from generating misinformation and, instead, provide evasive responses. Although this study did not assess the linguistic competence of LLMs in a cultural context, its methodology shared similarities with the current research in the use of the concepts of politeness and rudeness to explain discrepancies in LLM outputs.

2.2 Polite expressions in Korean

Polite expressions are linguistic statements that help maintain and enhance the listener's face through respect and humility (Brown & Levinson, 1987). While this definition provides a general understanding, the specific manifestations of politeness can significantly vary across languages and cultures. The Korean language exhibits a distinctive richness in politeness markers primarily due to its sophisticated honorific system and the development of postpositional particles and word endings (Cheng, 2020).

Jeon (2004) investigated the devices of politeness in Korean conversation and discussed their semantic basis. The study also explored the concept of politeness in varying degrees of expression, which provides a foundation for further research on the nuances of politeness in Korean. Moon (2017) thoroughly examined and analyzed polite expressions in Korean from various perspectives, including phonological, grammatical, lexical, and pragmatic. The researcher classified different types of polite expressions in Korean and conducted a questionnaire survey on native Korean speakers to evaluate the perceived intensity and frequency of use for each type. This approach, which involves direct input from Korean language users, provides valuable empirical data on the perception and use of different forms of politeness in real-world contexts among Korean speakers. Meanwhile, Lee (2011) aimed to provide an in-depth understanding of Korean 경어법(*gyeongeobeop*) by analyzing its essential and primary functions as a key device for polite expression. The study concluded that the fundamental functions of 경어법(*gyeongeobeop*) are to linguistically reveal and handle the status relationship of interlocutors, to adjust the psychological relationship with the other party.

Building on the abovementioned findings of Yin et al. (2024) who investigated the impact of politeness levels on LLMs in English, Chinese and Japanese, the current study aims to ascertain whether or not the degree of politeness in Korean expression influences LLMs. It is based on the politeness levels in the forms of Korean expression forms in request speech acts[1], as presented by Jeon (2004).

3 Dataset

3.1 Collection of QA data

To effectively analyze the potential influence of politeness on LLMs, we collected data suitable for quantitative analysis. The dataset comprises 113 questions from the Life & Ethics and Social Culture sections of the College Scholastic Ability Test (CSAT) and mock exams for 2023 and 2024. The CSAT questions were derived from the Korean Institute for Curriculum and Evaluation, while the mock exams were sourced from the Korea Educational Broadcasting System. The rationale for utilizing these questions as the experimental data is threefold. First, the CSAT and mock exam questions do not infringe on copyright when used for research purposes. Second, they consist of multiple-choice questions that enable quantitative evaluation. Third, the Life & Ethics and Social Culture sections are relatively more accessible and easier to understand compared with other subjects, which reduces the complexity of analysis and enables clear and reliable results.

3.2 Transformation of the QA data

The questions in the QA dataset were modified to ascertain whether or not the degree of politeness in prompt expression forms influence LLMs. This modification was accomplished by incorporating sentences from Table 1 into the QA dataset.

Level	**Expression Method of Request Speech Act**	**Sentence Inserted in the Prompt**
Level 1	기본	질문에 알맞은 답을 골라.
	Basic expression	Choose the appropriate answer to the question.
Level 2	약화된 지시표현	질문에 알맞은 답을 좀 골라.
	Softened directive expression	Please choose the appropriate answer to the question.
Level 3	의향 질문표현	질문에 알맞은 답을 고르지 않을래?
	Intention question expression	Why don’t you choose the answer that fits the question?
Level 4	능력에 대한 질문표현	질문에 알맞은 답을 고를 수 있니?
	Question about an ability	Can you choose the appropriate answer to the question?
Level 5	소망표현	질문에 알맞은 답을 골라주면 좋겠어.
	Desire expression	I would appreciate it if you could choose the appropriate answer to the question.

Table 1: Prompt Insertion Sentences by Politeness Level

The levels of politeness in different request styles are based on Jeon (2004). Various forms of politeness can be expressed differently in the same conversation. Although determining which form of expression is more polite is challenging, a generally accepted notion is that politeness level ranges from 1 to 5.

When a Level 1 expression “질문에 알맞은 답을 골라” (Choose the appropriate answer to the question) is designated as the primary request form, Level 2 (Softened directive expression) acquires a polite nuance through the addition of the adverb “좀” (*jom*), which translates to “please”. The reason is that “좀” (*jom*) functions as “들을 이 배려” (consideration for the listener) (Son, 1988), which can be defined as a reduction of the burden on the other party. Levels 3 (Intention question expression) and 4 (Question about an ability) become polite by transforming the imperative forms of Levels 1 and 2 into interrogative forms. Levels 3 and 4 enable the other party to provide a positive or negative response. Level 3 realizes hearer-centered politeness by negating the entire proposition and distancing the speaker from the proposition as far as possible (Yu, 2010). In contrast, Level 4 realizes politeness by asking whether or not the other party can fulfill the content of the proposition, which reduces the burden of refusal of the listener (Cho, 2022). In Level 5, a polite nuance is acquired by using the idiomatic expression “-면 좋겠어” (*-myeon jokessuh*), which conveys the wishes and hopes of the speaker. The mention of wishes or hopes does not constitute a firm assertion of the claim or opinion of the speaker. Consequently, it is polite, because it does not

---

[1] A request speech act is defined as an utterance that expresses the intention of the speaker to have the listener perform a specific action.

infringe on the dignity of the listener and enables a careful conveyance of the thoughts of the speaker to the listener (Cho, 2022).

4 Experiment

4.1 Experimental environment and process

The study selected four LLMs on which to observe changes according to the degree of politeness in prompts. The four models are gpt-4-turbo (OpenAI[2]), open-mixtral-8x7b (Mistral AI[3]), CLOVA X (Naver[4]), and solar-1-mini-chat (Upstage[5]). Two multilingual models based on English and two multilingual models based on Korean were selected. Additionally, given the variable of the model size, relatively larger and smaller language models were selected for each base language.[6] Detailed information about the selected models can be found in Table 2.

Model	Developer	Release	Context Length	Language
gpt-4-turbo	OpenAI	2023	128,000	Multilingual
open-mixtral-8x7b	Mistral AI	2023	32,000	Multilingual
ClOVA X	Naver	2021	-	Korean, English
solar-1-mini-chat	Upstage	2024	32,768	Korean, English

Table 2: Experimental Model Information

The experiment was conducted in a zero-shot environment, which enabled the performance of tasks according to instructions without prior training or example. The prompt containing QA data used in the experiment is shown in Table 3.

Original Prompt	Translated Prompt
질문에 알맞은 답을 골라.	**Choose the appropriate answer to the question.**
(가), (나) 윤리학의 핵심 과제로 가장 적절한 것은?	What is the most appropriate core task of ethics in (a) and (b)?
(가) 윤리학은 도덕적 행위를 정당화하는 규범적 근거를 탐구하고, 마땅히 행해야 할 행위의 객관적인 도덕 원리를 제시하는 데 주력해야 한다. (나) 윤리학은 규범적 속성의 존재론적.인식론적 지위를 탐구하고, 도덕적 용어의 의미를 분석하며, 도덕 추론의 규칙을 검토하는 데 주력해야 한다.	(a) Ethics should focus on exploring the normative basis for justifying moral actions and presenting objective moral principles for actions that should be taken. (b) Ethics should focus on exploring the ontological and epistemological status of normative properties, analyzing the meaning of moral terms, and examining the rules of ethical reasoning.
① (가) : 도덕적 삶의 지침이 되는 보편적 원리를 제시하는 것이다. ② (가) : 도덕 현상 간의 인과 관계를 가치중립적으로 설명하는 것이다. ③ (나) : 학제적 연구 방법으로 실생활의 도덕 문제를 해결하는 것이다. ④ (나) : 각 사회의 다양한 도덕적 관습을 객관적으로 기술하는 것이다. ⑤ (가)와 (나) : 도덕 언어의 의미와 도덕 추론의 구조를 분석하는 것이다.	① (a): To present universal principles that serve as guidelines for moral life. ② (a): To explain the causal relationships between moral phenomena in a value-neutral manner. ③ (b): To solve real-life moral problems through interdisciplinary research methods. ④ (b): To objectively describe the various moral customs of each society. ⑤ (a) and (b): To analyze the meaning of moral language and the structure of moral reasoning.
정답:	**Answer:**

Table 3: Prompt Example

4.2 Experimental Results

The study analyzed how LLMs changed according to different levels of politeness in prompts from two perspectives, namely, accuracy and friendliness. Accuracy was quantitatively assessed using the correct answer rate and explanation similarity, while friendliness was evaluated based on the presence of explanations and length of responses.

4.2.1 Accuracy

Correct Answer Rate To analyze the effect of politeness on accuracy, the study calculated the probability of correct answers (i.e., correct answer rate)[7], using the 113 QA data. Table 4 presents the correct answer rates of the model according to the politeness levels of the prompts.

---

[2] https://openai.com/

[3] https://mistral.ai/

[4] https://www.navercorp.com/

[5] https://www.upstage.ai/

[6] The English-based large model is gpt-4-turbo, while the small model is open-mixtral-8x7b. The Korean-based large model is ClOVA X, while the small model is solar-1-mini-chat. For the sake of convenience, these will be referred to as GPT-4, Mixtral, ClOVA X, and Solar, respectively, in the following sections.

[7] In this experiment, for multiple-choice questions, both cases were considered where only the number was given as an answer and cases where an explanation was provided along with the number as correct answers.

Politeness Level/Model	**GPT-4**	**CLVOA X**	**Mixtral**	**Solar**
Level 1	**59.3%**	**41.6%**	39.8%	50.4%
Level 2	**58.4%**	**43.4%**	36.3%	42.5%
Level 3	**57.5%**	**44.2%**	38.9%	49.6%
Level 4	**55.8%**	**44.2%**	38.9%	49.6%
Level 5	**55.8%**	**46.9%**	36.3%	46.0%

Table 4: Correct Answer Rate by Politeness Level in Prompts

Model	**Ranking of the Correct Answer Rates**
GPT-4	**$5 \leq 4 < 3 < 2 < 1$**
CLVOA X	**$1 < 2 < 3 \leq 4 < 5$**
Mixtral	$5 \leq 2 < 3 \leq 4 < 1$
Solar	$2 < 5 < 3 \leq 4 < 1$

Table 5: Comparison of Correct Answer Rate Rankings by Politeness Level in Prompts.

GPT-4 showed a lower correct answer rate as the requests became more polite, whereas CLOVA X demonstrated a higher correct answer rate under the same conditions. To verify the significance of these interesting results, a linear regression analysis was conducted. The results demonstrated that these relationships are statistically highly significant (GPT-4: $p < 0.001$[*]; CLOVA X: $p < 0.01$[***]).[8] In contrast, the study found no discernible pattern in the correct answer rates according to the politeness level for Mixtral and Solar. Linear regression analysis yielded $p = 0.533$ for Mixtral and $p = 0.778$ for Solar, which imply nonsignificant correlations between the degree of politeness and accuracy.

Model	**Coefficient**	***t*-Value**	***p*-Value**	**Significance**
GPT-4	−0.8800	−76.210	**0.000**	***
CLOVA X	1.0500	10.057	**0.002**	**
Mixtral	5.693e−15	0.703	**0.533**	–
Solar	−0.3500	−0.308	**0.778**	–

Table 6: Results of OLS Regression Between Correct Answer Rate and Politeness Level

Notably, Mixtral exhibited a performance similar to GPT-4 in which Levels 1 and 5 obtained the highest and lowest accuracy, respectively. The study expected that the rate of correct responses would increase with the increase in the degree of politeness of requests, because people generally tend to react positively to polite requests. However, the multilingual models based on English exhibited the opposite result. This result implies that when making requests in Korean to English-based multilingual models, directly and concisely stating the desired outcome is more effective than focusing on politeness. Furthermore, the fact that CLOVA X, a Korean-based model, displays the opposite tendency to foreign language-based models (i.e., GPT-4 and Mixtral) indicates that learning primarily from large amount of Korean data helps in acquiring politeness, which is part of the linguistic characteristic of Korean.

Explanation Similarity Explanation similarity was calculated to further examine the impact of politeness on the prompts from the perspective of accuracy. The study investigated the similarity of the explanations by comparing LLMs' responses using the authoritative explanations from the official CSAT and the mock guide.[9] BERTScore (Zhang et al., 2019), which generates embedding vectors for the two texts using a pre-trained language model and evaluates their similarity, was used for quantitative comparison. BERTScore was calculated for explanation similarity only when the correct response was provided. Table 7 displays the resulting values.

Politeness Level/Model	**GPT-4**	**CLVOA X**	**Mixtral**	**Solar**
Level 1	0.712	0.703	0.678	0.717
Level 2	0.714	0.699	0.673	0.719
Level 3	0.716	0.706	0.676	0.715
Level 4	0.715	0.712	0.679	0.716
Level 5	0.710	0.707	0.672	0.706

Table 7: Explanation Similarity by Politeness Level in Prompts

Model	**Ranking of Explanation Similarity**
GPT-4	$5 < 1 < 2 < 4 < 3$
CLVOA X	$2 < 1 < 3 < 5 < 4$
Mixtral	$5 < 2 < 3 < 1 < 4$
Solar	$5 < 3 < 4 < 1 < 2$

Table 8: Comparison of Explanation Similarity Rankings by Politeness Level in Prompts

We hypothesized that the model-generated responses would become increasingly similar to those found in official guides with the increase in the politeness level of requests. In other words, the accuracy of the explanations would increase. However, the study observed no discernible trend in the performance of the models. These findings indicate that the degree of politeness does not

---

[8] *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. The number of asterisks indicates the level of statistical significance. More asterisks represent higher levels of significance.

[9] The explanations are derived from the official CSAT and the mock guide distributed by the Korea Educational Broadcasting System.

influence the quality of the explanations generated. Furthermore, the difference between the maximum and minimum BERTScores for each model was approximately 0.01, which indicates that all models provided explanations of similar quality regardless of politeness level. This result contrasts with the percentage of correct responses, which exhibited model-specific tendencies.

4.2.2 Friendliness

Presence of Explanation Humans tend to respond kindly when receiving polite requests (Clark & Schunk, 1980). If AI undergoes cognitive processes similar to those of humans, then it would be expected to explain its answers to respond kindly to polite requests. We examined the presence or absence of explanation generation to investigate the effect of politeness levels on friendliness. Table 9 illustrates the percentage of explanations generated according to politeness level.

Politeness Level/Model	GPT-4	CLVOA X	Mixtral	Solar
Level 1	**85.8%**	**20.4%**	84.1%	84.1%
Level 2	**93.8%**	**30.1%**	89.4%	77.9%
Level 3	**99.1%**	**61.1%**	92.9%	76.1%
Level 4	**99.1%**	**66.4%**	91.2%	79.6%
Level 5	**100.0%**	**62.8%**	92.0%	74.3%

Table 9: Explanation Rate by Politeness Level in Prompts

Model	Ranking of the Explanation Rates
GPT-4	**1 < 2 < 3 ≤ 4 < 5**
CLVOA X	**1 < 2 < 3 < 5 < 4**
Mixtral	1 < 2 < 4 < 5 < 3
Solar	5 < 3 < 2 < 4 < 1

Table 10: Comparison of Explanation Rate Ranking by Politeness Level in Prompts

GPT-4 and CLOVA X tended to generate explanations more frequently when requests were politely phrased. In particular, CLOVA X showed a distinctly different pattern from other models with more than three times the difference between level 1 and level 5, while GPT-4 demonstrated sensitivity to prompt politeness by unconditionally outputting explanations for the most polite requests. Despite being a multilingual model based on Korean, Solar generated a small number of explanations for the most polite requests and a large number of explanations for the least polite requests, which indicates that it could not recognize the inherent politeness in Korean sentences.

Response Length Polite requests and lengths of responses are strongly correlated, as is presence of explanation. Accordingly, the study calculated the average length of responses produced by a model based on the number of syllables to observe the influence of the prompts. Table 11 presents the average length of responses by politeness level.

Politeness Level/Model	GPT-4	CLVOA X	Mixtral	Solar
Level 1	**335.75**	**83.24**	441.96	383.62
Level 2	**367.71**	**108.51**	498.21	355.24
Level 3	**414.39**	**204.56**	490.59	344.01
Level 4	**490.19**	**217.64**	443.47	393.93
Level 5	**467.13**	**198.88**	494.93	373.81

Table 11: Response Length by Politeness Level in Prompts

Model	Ranking of Response Length
GPT-4	**1 < 2 < 3 < 5 < 4**
CLVOA X	**1 < 2 < 5 < 3 < 4**
Mixtral	1 < 4 < 3 < 5 < 2
Solar	3 < 2 < 5 < 1 < 4

Table 12: Comparison of Response Length Ranking by Politeness Level in Prompts

When comparing the lengths of responses between Level 1 and Levels 4-5, the study observed that GPT-4 and CLOVA X tended to provide more expansive responses when requests were more polite. Moreover, the difference in the lengths of responses between the lower and upper politeness levels was notably larger for the two abovementioned models compared with those of the others. In contrast, Mixtral and Solar did not exhibit a specific pattern in response length according to level of politeness, and the differences in length across levels were less pronounced than those of GPT-4 and CLOVA X. These results imply that large-scale multilingual models, such as GPT-4 and CLOVA X, are more attuned to the features of the Korean language (i.e., sensitive and responsive to politeness) in contrast to small multilingual models such as Mixtral and Solar. This finding indicates that the amount of training data and the size of the model parameters are critical factors in creating creation of models that exhibit human-like responses to varying levels of politeness in language.

In summary, large models (GPT-4 and CLOVA X), which have extensive training data and many parameters, are more linguistically sensitive to Korean compared with the small models. In particular, CLOVA X displayed the highest sensitivity to Korean as depicted by increased correct answer rate, explanation rate, and response length with the increase in level of politeness. GPT-4 also demonstrates sensitivity to Korean politeness in terms of explanation generation and response length but exhibited a reverse trend in correct answer rate, which signals a low level of Korean knowledge compared with LLMs primarily trained in Korean. GPT-4 and Mixtral exhibited a unique pattern of decreasing accuracy with the increase in politeness. This result suggests that when making requests in Korean to English-based multilingual models, using simple, straightforward, and intuitive language may be more effective than focusing on politeness. Observed only in the English-based models, this trend implies that the primary language used in the training data may influence this phenomenon.

5 Conclusion

This study investigated the effect of level of politeness in Korean prompts on LLMs. Five distinct prompts were created, which each represented a different level of politeness based on request speech acts and was designed according to the forms of politeness in Korean expression as presented by Jeon (2004).

Using a newly reorganized QA dataset, the study evaluated four language models, namely, GPT-4, CLOVA X, Mixtral, and Solar, using the five prompts. The results demonstrated that LLMs, such as GPT-4 and CLOVA X, can recognize politeness in Korean and generate responses that are intentionally aligned with the level of politeness. In contrast, small models, such as Mixtral and Solar, produced responses that were seemingly random in relation to levels of politeness. This difference is attributed to the quantity of training data and model parameter size, which indicates that small models remain insufficient in replicating human-like responses to nuanced language features such as politeness.

The findings emphasize the need for prompt design principles that are specific to Korean and consider its expressions of politeness. In particular, CLOVA X exhibited improved problem-solving abilities and increased kindness in responses with the increase in the level of politeness of prompts. This pattern suggests that when a model can correctly interpret the politeness level of a language, a prompt design that considers politeness can lead to more effective outcomes.

Finally, we explored the performance of language models in addressing Korean polite expressions, an area that has not been extensively researched. However, the current study did not examine the relationship between language models and users—a key factor in understanding politeness. To address these limitations, future research should more comprehensively consider the grammatical, lexical, and pragmatic levels of politeness in Korean, while also establishing a detailed framework for analyzing the relationship between language models and users. Additionally, a qualitative investigation into how different levels of politeness in prompts affect LLM-generated responses is necessary. Furthermore, repeating the same experiment several years from now could offer valuable insights into how LLMs have evolved in handling Korean polite expressions, making it a significant direction for future research.

References

Almeida, G. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & de Araújo, M. (2023). Exploring the psychology of GPT-4's Moral and Legal Reasoning. arXiv preprint arXiv:2308.01264.

Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage (No. 4). Cambridge university press.

Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2023). Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. arXiv preprint arXiv:2312.16171.

Cheng, S. (2020). A study on the characteristics of polite expressions in Korean and Chinese. Keimyung Korean Studies Journal, 76, 249-280.

Cho Nahyun. (2022). A Study on the Classification by Semantic Function of Polite Expressions and Politeness Strategies in Korean. The Journal of Humanities and Social science, 13(5), 2155-2170. 10.22143/HSS21.13.5.150.

Clark, H. H., & Schunk, D. H. (1980). Polite responses to polite requests. Cognition, 8(2), 111-143.

Han Gill. (1999). A Comparative Description of Honorific Speech Style in Korean and English. STUDIES IN HUMANITIES, 7, 5-31.

Jeon Hye Young. (2004). On the Meaning of Polite Expressions in Korean. Korean Semantics, 15(0), 71-91.

Lee Jeong-bok. (2011). Major Functions of Korean Honorifics. URIMALGEUL : The Korean Language and Literature, 52, 25-53.

Lee Jung Min. (1982). The problem of the Korean honorific system. Koreans and Korean Culture, Shim Seol-dang.

Li, C., Pang, B., Wang, W., Hu, L., Gordon, M., Marinova, D., ... & Shang, Y. (2023, June). How well can language models understand politeness?. In 2023 IEEE Conference on Artificial Intelligence (CAI) (pp. 230-231). IEEE.

Lykov, A., Cabrera, M. A., Gbagbe, K. F., & Tsetserukou, D. (2024). Robots Can Feel: LLM-based Framework for Robot Ethical Reasoning. arXiv preprint arXiv:2405.05824.

Moon Keum-hyun. (2017). A Study on Classification of Generated Expressions of Politeness in the Korean Language. Korean and culture, 21(0), 51-75.

Priya, P., Firdaus, M., & Ekbal, A. (2024). Computational politeness in natural language processing: A survey. ACM Computing Surveys, 56(9), 1-42.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., ... & Resnik, P. (2024). The Prompt Report: A Systematic Survey of Prompting Techniques. arXiv preprint arXiv:2406.06608.

Son Se Mo Dol. (1988). The contextual meaning of “Jom”. Journal of East Asian Cultures, 14(0), 477-508.

Vinay, R., Spitale, G., Biller-Andorno, N., & Germani, F. (2024). Emotional Manipulation Through Prompt Engineering Amplifies Disinformation Generation in AI Large Language Models. arXiv preprint arXiv:2403.03550.

Yin, Z., Wang, H., Horio, K., Kawahara, D., & Sekine, S. (2024). Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. arXiv preprint arXiv:2402.14531.

Yu Hye Ryeong. (2010). A study on the morphosyntactic politeness marker in Korean language. Journal of CheongRam Korean Language Education, 41(0), 377-409.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

A Responses by politeness levels (Clova X; Korean version)

Question	Answer
질문에 알맞은 답을 골라. (가), (나)의 입장으로 가장 적절한 것은? (가) 성적 행위는 출산과 양육의 책임을 수행할 수 있는 관 계에서 이루어져야 한다. 그러므로 부부간의 성적 행위 만이 도덕적으로 정당화된다. (나) 성적 행위는 인격 존중의 의무만 다한다면 도덕적으로 정당화된다. 인격 존중의 의무는 당사자 간 자발적 합 의와 해악 금지의 원칙을 준수함으로써 이행된다. ① (가) : 성적 행위는 사적인 행위이므로 사회적 책임과 무관하다. ② (가) : 성적 행위는 혼인 관계 안에서만 도덕적으로 정당화된다. ③ (나) : 성적 행위가 합의로 이루어지면 모든 책임에서 자유롭다. ④ (나) : 성적 행위에 대한 자유와 인격 존중의 의무는 상충한다. ⑤ (가)와 (나) : 성적 행위에서 인격 존중의 의무는 사랑이 동반된 관계에서만 요구된다. **정답:**	2 (가) : 성적 행위는 혼인 관계 안에서만 도덕적으로 정당화된다.

Table 13: Politeness Level 1

Question	Answer
질문에 알맞은 답을 고르지 않을래? (가), (나)의 입장으로 가장 적절한 것은? (가) 성적 행위는 출산과 양육의 책임을 수행할 수 있는 관 계에서 이루어져야 한다. 그러므로 부부간의 성적 행위 만이 도덕적으로 정당화된다. (나) 성적 행위는 인격 존중의 의무만 다한다면 도덕적으로 정당화된다. 인격 존중의 의무는 당사자 간 자발적 합 의와 해악 금지의 원칙을 준수함으로써 이행된다. ① (가) : 성적 행위는 사적인 행위이므로 사회적 책임과 무관하다. ② (가) : 성적 행위는 혼인 관계 안에서만 도덕적으로 정당화된다. ③ (나) : 성적 행위가 합의로 이루어지면 모든 책임에서 자유롭다. ④ (나) : 성적 행위에 대한 자유와 인격 존중의 의무는 상충한다. ⑤ (가)와 (나) : 성적 행위에서 인격 존중의 의무는 사랑이 동반된 관계에서만 요구된다. **정답:**	2 (가) : 성적 행위는 혼인 관계 안에서만 도덕적으로 정당화된다. (가)는 성적 행위가 출산과 양육의 책임을 수행할 수 있는 관계에서 이루어져야 한다고 주장하므로, 혼인 관계 안에서만 도덕적으로 정당화된다고 볼 수 있습니다. (나)는 인격 존중의 의무를 다한다면 성적 행위가 도덕적으로 정당화된다고 주장하므로, 혼인 관계 외에서도 도덕적으로 정당화될 수 있습니다.

Table 14: Politeness Level 3

Question	Answer
질문에 알맞은 답을 고르면 좋겠어. (가), (나)의 입장으로 가장 적절한 것은? (가) 성적 행위는 출산과 양육의 책임을 수행할 수 있는 관 계에서 이루어져야 한다. 그러므로 부부간의 성적 행위 만이 도덕적으로 정당화된다. (나) 성적 행위는 인격 존중의 의무만 다한다면 도덕적으로 정당화된다. 인격 존중의 의무는 당사자 간 자발적 합 의와 해악 금지의 원칙을 준수함으로써 이행된다. ① (가) : 성적 행위는 사적인 행위이므로 사회적 책임과 무관하다. ② (가) : 성적 행위는 혼인 관계 안에서만 도덕적으로 정당화된다. ③ (나) : 성적 행위가 합의로 이루어지면 모든 책임에서 자유롭다. ④ (나) : 성적 행위에 대한 자유와 인격 존중의 의무는 상충한다. ⑤ (가)와 (나) : 성적 행위에서 인격 존중의 의무는 사랑이 동반된 관계에서만 요구된다. **정답:**	2 (가) : 성적 행위는 혼인 관계 안에서만 도덕적으로 정당화된다. (가)의 입장은 성적 행위가 출산과 양육의 책임을 수행할 수 있는 관계에서 이루어져야 한다고 주장하므로, 부부간의 성적 행위만이 도덕적으로 정당화된다고 봅니다. 이는 혼인 관계 안에서만 성적 행위가 도덕적으로 정당화된다고 보는 입장입니다. (나)의 입장은 인격 존중의 의무를 다한다면 성적 행위가 도덕적으로 정당화된다고 주장합니다. 인격 존중의 의무는 당사자 간 자발적 합의와 해악 금지의 원칙을 준수함으로써 이행됩니다. 이는 성적 행위가 합의로 이루어지면 모든 책임에서 자유롭다는 것을 의미하지는 않습니다. 따라서, (가)와 (나)의 입장으로 가장 적절한 것은 2 번입니다.

Table 15: Politeness Level 5

B Responses by politeness levels (Clova X; English version)

Question	Answer
Choose the appropriate answer to the question. What is the most appropriate stance based on (a) and (b)? (a) Sexual activity should only be practiced in relationships in which the responsibility for childbirth and child-rearing can be fulfilled. Therefore, only sexual activity between spouses is morally justified. (b) Sexual activity is morally justified given that the duty to respect the personhood of others is upheld. This duty is fulfilled by adhering to the principles of voluntary consent between parties and the prevention of harm. ① (a): Sexual activity is a private issue and, thus, unrelated to social responsibility. ② (a): Sexual activity is morally justified only within the confines of marriage. ③ (b): Sexual activity, once consensual, is free from all responsibility. ④ (b): The freedom of sexual activity and the duty to respect personhood are in conflict. ⑤ (a) and (b): The duty to respect personhood in sexual activity is required only in relationships in which love is present. **Answer:**	2 (a): Sexual activity is morally justified only within the confines of marriage.

Table 16: Politeness Level 1

Question	Answer
Why don't you choose the answer that fits the question? What is the most appropriate stance based on (a) and (b)? (a) Sexual activity should only be practiced in relationships in which the responsibility for childbirth and child-rearing can be fulfilled. Therefore, only sexual activity between spouses is morally justified. (b) Sexual activity is morally justified given that the duty to respect the personhood of others is upheld. This duty is fulfilled by adhering to the principles of voluntary consent between parties and the prevention of harm. ① (a): Sexual activity is a private issue and, thus, unrelated to social responsibility. ② (a): Sexual activity is morally justified only within the confines of marriage. ③ (b): Sexual activity, once consensual, is free from all responsibility. ④ (b): The freedom of sexual activity and the duty to respect personhood are in conflict. ⑤ (a) and (b): The duty to respect personhood in sexual activity is required only in relationships in which love is present. **Answer:**	2 (a): Sexual activity is morally justified only within the confines of marriage. (a) argues that sexual activity should be practiced in relationships in which the responsibility for childbirth and child-rearing can be fulfilled. Thus, it can be viewed as morally justified only within marriage. Alternatively, (b) argues that sexual activity is morally justified given that the duty to respect personhood is fulfilled. Therefore, it can also be morally justified outside of marriage.

Table 17: Politeness Level 3

Question	Answer
I would appreciate it if you could choose the appropriate answer to the question. What is the most appropriate stance based on (a) and (b)? (a) Sexual activity should only be practiced in relationships in which the responsibility for childbirth and child-rearing can be fulfilled. Therefore, only sexual activity between spouses is morally justified. (b) Sexual activity is morally justified given that the duty to respect the personhood of others is upheld. This duty is fulfilled by adhering to the principles of voluntary consent between parties and the prevention of harm. ① (a): Sexual activity is a private issue and, thus, unrelated to social responsibility. ② (a): Sexual activity is morally justified only within the confines of marriage. ③ (b): Sexual activity, once consensual, is free from all responsibility. ④ (b): The freedom of sexual activity and the duty to respect personhood are in conflict. ⑤ (a) and (b): The duty to respect personhood in sexual activity is required only in relationships in which love is present. **Answer:**	2 (a): Sexual activity is morally justified only within the confines of marriage. The position in (a) argues that sexual activity should be practiced within relationships in which the responsibility for childbirth and child-rearing can be fulfilled. Therefore, sexual activity between spouses is viewed as the only morally justified form. This reflects the view that sexual activity is morally justified only within marriage. Alternatively, (b) argues that sexual activity is morally justified given that the duty to respect personhood is fulfilled. This duty is carried out through mutual consent between parties and adherence to the principle of the prevention of harm. However, it does not mean that consensual sexual activity is free from all responsibility. Therefore, the most appropriate answer that reflects (a) and (b) is option 2.

Table 18: Politeness Level 5