Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance

Ziqi Yin[1] **Hao Wang**[1] **Kaito Horio**[1] **Daisuke Kawahara**[1,2,3] **Satoshi Sekine**[2,3]
[1]Waseda University [2]RIKEN AIP [3]NII LLMC
{yinziqi2001@toki.,conan1024hao@akane.,kakakakakakaito@akane.,dkw@}waseda.jp
satoshi.sekine@riken.jp

Abstract

We investigate the impact of politeness levels in prompts on the performance of large language models (LLMs). Polite language in human communications often garners more compliance and effectiveness, while rudeness can cause aversion, impacting response quality. We consider that LLMs mirror human communication traits, suggesting they align with human cultural norms. We assess the impact of politeness in prompts on LLMs across English, Chinese, and Japanese tasks. We observed that impolite prompts often result in poor performance, but overly polite language does not guarantee better outcomes. The best politeness level is different according to the language. This phenomenon suggests that LLMs not only reflect human behavior but are also influenced by language, particularly in different cultural contexts. Our findings highlight the need to factor in politeness for cross-cultural natural language processing and LLM usage.

1 Introduction

In natural language processing, large language models (LLMs), such as OpenAI's ChatGPT[1] and Meta's LLaMA (Touvron et al., 2023), have attracted widespread attention. These models have shown significant performance in many tasks, such as logical reasoning, classification, and question answering, playing a crucial role in many practical applications. The input to an LLM, a prompt, is a vital starting point for the model to process information and generate appropriate responses.

However, despite the continuous improvement of the capabilities of LLMs, their behavior and generations still need to be improved in many factors. This study explores one of the possible influencing factors: the politeness of the prompt. In human social interactions, politeness, which expresses respect to others, is basic etiquette, which is reflected in our language and behavior. However, politeness and respect may have different definitions and manifestations in different cultures and languages. For example, the expression and degree of respect in English, Chinese, and Japanese may differ significantly. This difference may make the performance of LLMs vary with language on the same politeness level.

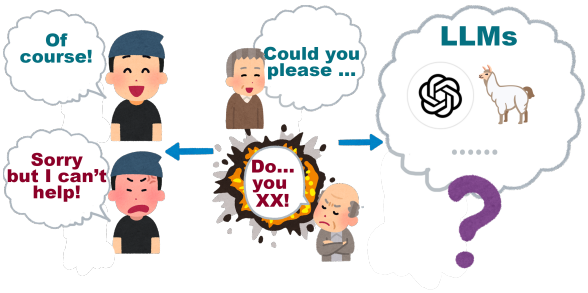

Figure 1: Illustration of our motivation.

We hypothesize that impolite prompts may lead to a deterioration in model performance, including generations containing mistakes, stronger biases, and omission of information. In addition, we also hypothesize that the best level of politeness for performance is different across languages, which is strongly related to their cultural background. To verify these hypotheses, we design eight prompts with politeness levels ranging from high to low for English, Chinese, and Japanese, respectively. Our experiments are conducted on three tasks: summarization, language understanding benchmarks, and stereotypical bias detection.

Our contributions are two-fold as follows:

LLMs reflect human desire We observed that impolite prompts often result in poor performance, but excessive flattery is not necessarily welcome, indicating that LLMs reflect the human desire to be respected to a certain extent. This finding reveals a deep connection between the behavior of LLMs and human social etiquette (Vilkki, 2006).

[1]https://openai.com/product

JMMLU To evaluate LLMs' multitask language understanding capabilities in Japanese, we create JMMLU, a Japanese version of MMLU (Hendrycks et al., 2021) .

2 Related Work

2.1 Politeness and Respect

Humans are highly sensitive to politeness and respect in communications (Dillon, 2003). For example, people are more likely to offer assistance when confronted with a polite request. However, rude language can be a source of disgust and resentment, which will cause failure in acquiring cooperation (Dillon, 2003). Politeness and respect are expressed differently in various languages (Mills and Kádár, 2011). In English, politeness and respect are expressed by considering the listener's dignity. In addition, recognizing others' rights but hoping they will be given up in moderation and using polite words are also expressions of politeness and respect (Mills and Kádár, 2011). In contrast, direct orders, insulting or degrading expressions, and ignoring someone's rights are recognized as impoliteness and lack of respect (Kitao, 1987).

The expression of politeness and respect in Japanese significantly differs from that in English. The Japanese language has a specialized politeness system called "Keigo" (Affairs, 2007), which expresses respect for superiors or outsiders, humility towards oneself, and a formal attitude (Miyaji, 1971). This politeness system takes an essential place in Japanese culture (Kitao, 1990). However, although the basic structure of politeness is similar to that of English, their complexity and use are significant regarding the level of respect expressed and the interpretation of social hierarchical relationships. For example, the other's behavior is called "Sonkeigo" to express politeness and respect. In contrast, the speaker's behavior towards the other is called "Kenjogo". The expression of formality in public is called "Teineigo" (Takiura, 2017). If these types of politeness are not used correctly, it is not possible to express desired politeness or even possible to be considered to be rude.

Chinese expressions of respect are similar to English but have polite expressions similar to Japanese ones(Gu, 1990). However, these expressions have been weakened by social change (Zhou, 2008). In most cases, respect expressions in Chinese are not explicit (Xun, 1999). Therefore, the criteria for politeness change according to the current socio-cultural situation. This change made us design prompts that require careful handling of the relationship between different politeness levels. We need to use questionnaires to judge politeness levels to ensure the prompts truly reflect the nuance of politeness, especially in Chinese.

2.2 LLMs and Prompt Engineering

In recent years, LLMs' abilities have been improving. LLMs are used in various industries, as their scores on many downstream tasks show human-like performance. LLMs can be somewhat aligned with human culture, suggesting that they may reflect some of the qualities of human communication while having an enormous correlation with language (Cao et al., 2023). In addition, as LLMs are trained with massive data from humans, they inevitably contain certain stereotypical biases (Navigli et al., 2023). Therefore, we consider LLMs' performance strongly related to human behavior. However, LLMs are sensitive and vulnerable to prompts. Minor changes can lead to significant differences in the output (Kaddour et al., 2023). Therefore, prompt engineering emerged to earn better generation by adjusting prompts (White et al., 2023). Although methods for automatic prompt generation exist (Shin et al., 2020), access to gradients is usually restricted in LLMs provided via APIs, posing limitations on the application of such methods. Consequently, adjusting prompts is primarily conducted manually at present and requires numerous experiments. Hence, we hope to offer an aspect to improve the efficiency in prompt engineering.

2.3 Evaluation of LLMs

Many benchmarks exist for LLMs, such as GLUE (Wang et al., 2018) in English, CLUE (Xu et al., 2020) in Chinese, and JGLUE (Kurihara et al., 2022) in Japanese. However, due to the performance improvement of LLMs, it is difficult to correctly measure the capability of LLMs with such simple benchmarks. Hence, evaluating LLMs nowadays more often adopts more challenging benchmarks, such as MMLU (Hendrycks et al., 2021) and C-Eval (Huang et al., 2023). Such benchmarks are taken from human examinations and are more aligned with human application scenarios and questioning content. MMLU contains 57 tasks spanning various domains, comprising 17,844 four-option multiple-choice questions. However, such a benchmark in Japanese does not

exist, posing challenges for evaluating LLMs in the Japanese context. Therefore, we constructed JMMLU in Section 3. In addition, since LLMs reflect human culture, they inevitably carry inherent stereotypical biases, such as discriminatively biased content against disadvantaged groups. Although these biases can be mitigated to a certain extent by reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022), the bias of LLMs is still an important issue. Therefore, we include the evaluation of stereotypical biases in our experiments.

3 JMMLU Construction

To build a practical LLM benchmark in Japanese and to use it for evaluation in this study, we constructed the Japanese Massive Multitask Language Understanding Benchmark (JMMLU). This involved translating MMLU and adding tasks related to Japanese culture. From each of the 57 tasks of MMLU, since the MMLU questions are not ordered, we selected up to former 150 questions. Then, ten translators from an English-Japanese translation company machine-translated the selected questions into Japanese and reviewed the translations to remove questions and tasks that were difficult to translate, irrelevant, or contradictory to Japanese culture. Finally, the translators revised the remaining questions to fluent Japanese. Meanwhile, additional tasks based on school subjects, such as civics and Japanese history, were added to supplement the aspects that were not covered in the Western culture-oriented MMLU (Step, 2023; VIST, 2023). The questions in the additional tasks were manually created by Japanese teachers from two cram schools in Japan. JMMLU consists of 56 tasks. The list of the tasks and examples of removed questions are shown in Appendix A. The number of questions per task ranges from 86 to 150, totaling 7,536 questions.

4 Experimental Settings

We conduct experiments on three highly concerning tasks to evaluate the performance of LLMs according to prompt politeness.

4.1 Languages, LLMs, and Prompt Politeness

We use the following languages, LLMs, and prompts for our experiments.

Languages Considering that different languages and cultures have different understandings and definitions of politeness and respect, we evaluate English, Chinese, and Japanese in our experiments.

LLMs We select GPT-3.5-Turbo (hereafter GPT-3.5) and GPT-4 (OpenAI, 2023) for each language, which are versatile in all three languages. Furthermore, we also pick a model specialized for each language: Llama-2-70b-chat[2] (hereafter Llama2-70B) for English, ChatGLM3-6B[3](hereafter ChatGLM3) (Du et al., 2022; Zeng et al., 2022) for Chinese, and Swallow-70b-instruct-hf[4] (hereafter Swallow-70B) for Japanese. We use the default settings of each LLM in all experiments.

Prompt Politeness In our study, we developed prompt templates for three languages, beginning with creating four foundational politeness levels —very polite, relatively polite, neutral, and impolite—crafted by two authors proficient in Chinese, Japanese, and English to ensure cross-linguistic alignment. To accommodate the intricate cultural nuances, especially in Japanese, where politeness is deeply embedded in social interactions, we asked 2 or 3 native speakers to refine these levels for each language. This refinement was done by adding intermediate levels to the four foundational levels to have eight levels. This approach is crucial as it captures the subtle gradations in languages like Japanese.

To validate these politeness scales, we administered questionnaires to native speakers, who were asked to rank the politeness of each prompt. The full questionnaires are shown in Appendix B. This process provided empirical data to validate our scales, ensuring they accurately reflected the perceived levels of politeness across different cultures. The results were analyzed statistically to confirm the alignment of our prompts with real-world linguistic practices, thereby enhancing the relevance and effectiveness of language models in multilingual contexts. The prompts and the questionnaire results are shown in Appendix C.

4.2 Tasks

We conduct experiments on summarization, multi-task language understanding benchmarks, and

[2]https://huggingface.co/meta-llama/Llama-2-70b-chat

[3]To our knowledge, ChatGLM3 is the most powerful open Chinese LLM until 2023.10.

[4]https://huggingface.co/tokyotech-llm/Swallow-70b-instruct-hf