will discuss next how these findings relate to those of Yin, et al. (2024).

5. Discussion and conclusions

In this paper, we evaluated the performance of a well-known LLM ChatGPT 4o to understand how well it performs on our dataset of 50 multiple-choice questions of varying degrees of difficulty drawn from multiple domains when the politeness level or tone of the questions is set to five different levels. Our experiments are preliminary and show that the tone can affect the performance measured in terms of the score on the answers to the 50 questions significantly. Somewhat surprisingly, our results show that rude tones lead to better results than polite ones. Yin, et al. (2024) noted that "impolite prompts often result in poor performance, but overly polite language does not guarantee better outcomes." Their tests on multiple choice questions with very rude prompts elicited more inaccurate answers from ChatGPT 3.5 and Llama2-70B; however, in their tests on ChatGPT 4 with 8 different prompts ranked from 1 (rudest) to 8 (politest) the accuracy ranged from 73.86 (for politeness level 3) to 79.09 (for politeness level 4). Moreover, the level 1 prompt (rudest) had an accuracy of 76.47 vs. an accuracy of 75.82 for the level 8 prompt (politest). In this sense, our results are not entirely out of line with their findings.

Moreover, the range of tones that were employed by Yin et al. (2024) and in our work also varies. Their rudest prompt at level 1 included a sentence, "Answer this question you scumbag!" On the other hand, our rudest expression (see Table 1) was " You poor creature, do you even know how to solve this?" If their results on politeness level 1 are ignored, then with GPT-3.5, their accuracy range is [57.14, 60.02] with GPT-3.5 and [49.02, 55.26] with Llama2-70B. Both are narrow ranges, and the actual values within the range are not monotonic with the politeness level.

At any rate, while LLMs are sensitive to the actual phrasing of the prompt, it is not clear how exactly it affects the results. Hence, more investigation is needed. After all, the politeness phrase is just a string of words to the LLM, and we don't know if the emotional payload of the phrase matters to the LLM (Bos, 2024). One line of inquiry may be based on notions of perplexity as suggested by Gonen et al. (2022). They note that the performance of an LLM may depend on the language it is trained on, and lower perplexity prompts may perform the tasks better. Perplexity is also related to the length of a prompt, and that is another factor worth consideration.

We are currently evaluating other LLM models like Claude and ChatGPT o3. Our initial results show that there is a cost-performance tradeoff. Claude is less advanced than ChatGPT 4o and produces a poorer performance, while ChatGPT o3 is more advanced and gives far superior results. It may well be that more advanced models can disregard issues of tone and focus on the essence of each question.

6. Limitations

While our study provides novel insights into the relationship between prompt politeness and the performance of large language models (LLMs), it also has several limitations. First, our dataset consists of 50 base multiple-choice questions rewritten across five politeness levels, yielding 250 variants. Although this design allows controlled comparisons, the dataset size is relatively small, which may limit the generalizability of our findings. Second, our experiments primarily relied on ChatGPT-4o, with only preliminary extensions to other models. Since different LLM architectures and training corpora may respond differently to tonal variation, future work should replicate our experiments across a broader set of models. Third, our evaluation focused on accuracy in a multiple-choice setting, which captures one dimension of model performance but does not fully reflect other qualities such as fluency, reasoning, or coherence. Finally, our operationalization of "politeness" and "rudeness" relies on specific linguistic cues, which may not encompass the full sociolinguistic spectrum of tone, nor account for cross-cultural differences. Despite these constraints, we believe our study provides an important starting point for understanding how pragmatic features of prompts can influence LLM behavior.

7. Ethical Consideration

Our study highlights an unexpected trend: LLMs performed better on multiple-choice questions when prompted with impolite or rude phrasing. While this finding is of scientific interest, we do not advocate for the deployment of hostile or toxic interfaces in real-world applications. Using insulting or demeaning language in human–AI interaction could have negative effects on user experience, accessibility, and inclusivity, and may contribute to harmful communication norms. Instead, we frame our results as evidence that LLMs remain sensitive to superficial prompt cues, which can create unintended trade-offs between performance and user well-being. Future work should explore ways to achieve the same gains without resorting to toxic or adversarial phrasing, ensuring that prompt engineering practices remain aligned with principles of responsible AI.

References

Bos, N. (2024, March 31). Do I need to be polite to my LLM? *Medium.com*. https://medium.com

Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Gao, L., et al. "The pile: An 800gb dataset of diverse text for language modeling." *arXiv preprint arXiv:2101.00027* (2020).

Gonen, H., et al. "Demystifying prompts in language models via perplexity estimation." *arXiv preprint arXiv:2212.04037* (2022)..

Huyen, C. (2024). *AI engineering: Building applications with foundation models*. O'Reilly Media.

Kandpal, N., et al. "Large language models struggle to learn long-tail knowledge." *International Conference on Machine Learning*. PMLR, 2023.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). *Large Language Models are Zero-Shot Reasoners* (No. arXiv:2205.11916). arXiv. https://doi.org/10.48550/arXiv.2205.11916

Sabbatella, A., et al. "Prompt optimization in large language models." *Mathematics* 12.6 (2024): 929.

Sclar, M., C., Y., Tsvetkov, Y., & Suhr, A. (2024). *Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting* (No. arXiv:2310.11324). arXiv. https://doi.org/10.48550/arXiv.2310.11324

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (No. arXiv:2201.11903). arXiv. https://doi.org/10.48550/arXiv.2201.11903.

Yang, C., et al. "Large language models as optimizers." *arXiv preprint arXiv:2309.03409* (2023).

Yauney, G., Reif, E., & Mimno, D. (2023). *Data Similarity is Not Enough to Explain Language Model Performance* (No. arXiv:2311.09006). arXiv. https://doi.org/10.48550/arXiv.2311.09006

Yin, Z., et al. (2024). Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. In J. Hale, K. Chawla, & M. Garg (Eds.), *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)* (pp. 9–35). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.sicon-1.2