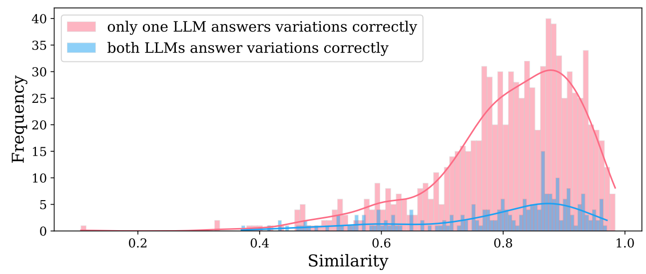

Fig. 2: Distribution of answerability of variations of the questions in `PromptSET` test sets which both LLMs failed to answer correctly.

shown in Figure 1 (b) and (c).We present the distribution of correctly and incorrectly *answered prompts* and in Figure 1(b), and *predicted responses* in Figure 1(c), based on similarity to the original prompt. Similarity is measured using the cosine similarity of the embedded representations of prompt-variation pairs, calculated with MiniLM, a model known for its strong performance in various NLP and IR tasks [37]. We observe that when a variation closely resembles the original prompt, it is more likely to generate both correct responses and accurate predictions of answerability. This suggests that the model may have encountered this data points before, indicating a strong bias toward its training data and reduced generalizability to less familiar or novel prompt formulations.

Impact of choice of LLM on variation answerability. We further explore whether prompt reformulation can enhance the effectiveness of an LLM. To investigate this, we first filter out questions from the `PromptSET` test set for which both LLMs, namely LLaMA and Mistral, failed to answer the original prompt correctly. Next, we examine the variations of these questions to see if an alternative prompt allows either LLM to provide a correct answer. The results are shown in Figure 2. For each sample in this figure, both LLMs failed to answer the original prompt correctly. However, in the red cases, at least one of the two LLMs succeeded in answering a variation correctly, while in the blue cases, both LLMs provided correct answers to the variation. This highlights the potential of prompt reformulation as a strategy. We conclude that `PromptSET` can serve as a valuable resource for prompt reformulation, helping transform an unanswerable prompt into an answerable one through LLM-driven reformulation.

4 Concluding Remarks

This paper investigates the sensitivity of LLMs to prompt variations by introducing the Prompt Sensitivity Prediction task and the `PromptSET` dataset, based on TriviaQA and HotpotQA. We generate variations of different questions and examine the sensitivity of various LLMs to these variations, all of which share the same underlying information need. Our benchmarking results reveal that existing methods do not fully capture the complexities of prompt sensitivity. These findings underscore the need for further research into prompt variation sensitivity, particularly in developing methods to help users generate more reliable prompts.

References

1. Arabzadeh, N., Bigdeli, A., Zihayat, M., Bagheri, E.: Query performance prediction through retrieval coherency. In: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43. pp. 193–200. Springer (2021)
2. Arabzadeh, N., Hamidi Rad, R., Khodabakhsh, M., Bagheri, E.: Noisy perturbations for estimating query difficulty in dense retrievers. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 3722–3727 (2023)
3. Arabzadeh, N., Huo, S., Mehta, N., Wu, Q., Wang, C., Awadallah, A.H., Clarke, C.L.A., Kiseleva, J.: Assessing and verifying task utility in LLM-powered applications. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21868–21888. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.1219, https://aclanthology.org/2024.emnlp-main.1219/
4. Arabzadeh, N., Meng, C., Aliannejadi, M., Bagheri, E.: Query performance prediction: From fundamentals to advanced techniques. In: European Conference on Information Retrieval. pp. 381–388. Springer (2024)
5. Arabzadeh, N., Meng, C., Aliannejadi, M., Bagheri, E.: Query performance prediction: Techniques and applications in modern information retrieval. In: Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region. pp. 291–294 (2024)
6. Arabzadeh, N., Seifikar, M., Clarke, C.L.: Unsupervised question clarity prediction through retrieved item coherency. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. pp. 3811–3816 (2022)
7. Arabzadeh, N., Zarrinkalam, F., Jovanovic, J., Al-Obeidat, F., Bagheri, E.: Neural embedding-based specificity metrics for pre-retrieval query performance prediction. Information Processing & Management **57**(4), 102248 (2020)
8. Arabzadeh, N., Zarrinkalam, F., Jovanovic, J., Bagheri, E.: Neural Embedding-Based Metrics for Pre-retrieval Query Performance Prediction, p. 78–85. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-45442-5_10, http://dx.doi.org/10.1007/978-3-030-45442-5_10
9. Arabzadeh, N., Zarrinkalam, F., Jovanovic, J., Bagheri, E.: Geometric estimation of specificity within embedding spaces. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. p. 2109–2112. CIKM '19, ACM (Nov 2019). https://doi.org/10.1145/3357384.3358152, http://dx.doi.org/10.1145/3357384.3358152
10. Bhargava, A., Witkowski, C., Shah, M., Thomson, M.W.: What's the magic word? a control theory of llm prompting. ArXiv **abs/2310.04444** (2023). https://doi.org/10.48550/arXiv.2310.04444
11. Bigdeli, A., Arabzadeh, N., Bagheri, E.: Learning to jointly transform and rank difficult queries. In: European Conference on Information Retrieval. pp. 40–48. Springer (2024)
12. Collins-Thompson, K., Bennett, P.N.: Predicting Query Performance via Classification, p. 140–152. Springer Berlin Heidelberg (2010). https://doi.org/10.1007/978-3-642-12275-0_15, http://dx.doi.org/10.1007/978-3-642-12275-0_15
13. Ebrahimi, S., Khodabakhsh, M., Arabzadeh, N., Bagheri, E.: Estimating query performance through rich contextualized query representations. In: European Conference on Information Retrieval. pp. 49–58. Springer (03 2024). https://doi.org/10.1007/978-3-031-56066-8_6

14. Faggioli, G., Ferro, N., Muntean, C.I., Perego, R., Tonellotto, N.: A geometric framework for query performance prediction in conversational search. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1355–1365. SIGIR '23, ACM (Jul 2023). https://doi.org/10.1145/3539618.3591625, http://dx.doi.org/10.1145/3539618.3591625
15. Feng, Z., Zhou, H., Zhu, Z., Qian, J., Mao, K.: Unveiling and manipulating prompt influence in large language models. ArXiv **abs/2405.11891** (2024), https://api.semanticscholar.org/CorpusID:269922034
16. Gasparetto, A., Marcuzzo, M., Zangari, A., Albarelli, A.: A survey on text classification algorithms: From text to predictions. Information **13**(2), 83 (2022)
17. Hambarde, K.A., Proença, H.: Information retrieval: Recent advances and beyond. IEEE Access **11**, 76581–76604 (2023)
18. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: CIKM (2008)
19. Hosseini, S.M., Arabzadeh, N., Zihayat, M., Bagheri, E.: Enhanced retrieval effectiveness through selective query generation. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. pp. 3792–3796 (2024)
20. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), https://arxiv.org/abs/2310.06825
21. Joshi, M., Choi, E., Weld, D., Zettlemoyer, L.: TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In: Barzilay, R., Kan, M.Y. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1601–1611. Association for Computational Linguistics, Vancouver, Canada (Jul 2017). https://doi.org/10.18653/v1/P17-1147, https://aclanthology.org/P17-1147
22. Kamalloo, E., Dziri, N., Clarke, C., Rafiei, D.: Evaluating open-domain question answering in the era of large language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.acl-long.307, http://dx.doi.org/10.18653/v1/2023.acl-long.307
23. Khodabakhsh, M., Zarrinkalam, F., Arabzadeh, N.: BertPE: A BERT-Based Pre-retrieval Estimator for Query Performance Prediction, p. 354–363. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-56063-7_27, http://dx.doi.org/10.1007/978-3-031-56063-7_27
24. Lo, L.S.: The art and science of prompt engineering: A new literacy in the information age. Internet Reference Services Quarterly **27**, 203 – 210 (2023). https://doi.org/10.1080/10875301.2023.2227621
25. Loya, M., Sinha, D., Futrell, R.: Exploring the sensitivity of llms' decision-making capabilities: Insights from prompt variations and hyperparameters. In: Findings of the Association for Computational Linguistics: EMNLP 2023. p. 3711–3716. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.findings-emnlp.241, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.241
26. Meng, C., Arabzadeh, N., Aliannejadi, M., de Rijke, M.: Query performance prediction: From ad-hoc to conversational search. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2583–2593. SIGIR '23, ACM (Jul 2023). https://doi.org/10.1145/3539618.3591919, http://dx.doi.org/10.1145/3539618.3591919