Krystian Ayala Due: April 8th, 2023

# "Are Prompt-based Models Clueless?" Summary

The long paper, "Are Prompt-based Models Clueless?", written by Pride Kavumba, Ryo Takahashi, and Yusuke Oda, who are affiliated with Tohoku University, RIKEN AIP, and LegalForce Research, mainly concerns testing the prompt-based models to ensure they learned in a more generalized way and not searching for special patterns based off a given data set. It was found by testing several different datasets that the prompt-based model did in fact use superficial cues to determine answers to given tasks. Kavumba divided the main question in to two separate ones, "Do prompt-based models exploit context superficial cues?" and "Do prompt-based models exploit contextless superficial cues?" After some more testing was done, it could be seen that the model used did exploit both contextual and not contextual cues given certain prompts. The next question that came to mind was determining the models' sensitivity to word order based on different training sets. The model does not have any real issues with this adjustment and the predictions are mostly the same. Which means that the model does not particularly rely on the meanings of given instances. Some additional related works are provided, and the results indicate that prompt-based models do exploit superficial cues. They do not generalize well on instances without superficial cues given by the used datasets. It was discovered that there were more superficial cues than known previously. Some prior works of the main author, Pride Kavumba include: "COPA-SSE: Semi-structured Explanations for the Commonsense Reasoning", "Learning to Learn to be Right for the Right Reasons", "When Choosing Plausible Alternatives, Clever Hans can be Clever", and "Improving Evidence Detection by Leveraging Warrants". The majority of these previous works pertain largely to understanding

how models use superficial cues in order to provide answers to certain tasks or problems. As well as why models tend to use these superficial cues. When it came to evaluating their work, the author would come up with several questions upon seeing the results of their initial problem. Like adjusting correct answers in a specific task to see if the model could still determine the right answer. Removing contextual cues in the answers was the main adjustment that was made to the problems / tasks. All the experiments were run 3 times with different random seeds, and they reported the average and standard deviation. It was determined that the model tested did “exploit contextless superficial cues” in particular on a larger training set but does not in smaller training sets. Kavumba then wondered if the prompt-based models were sensitive to a question’s meaning after analyzing these results. Further adjustments were made to the examples to continue testing this hypothesis. Pride Kavumba has received a total of 47 citations based on the number from Google Scholar. Their work is particularly important in understanding how a prompt-based model learns and its reasoning for its responses. Kavumba focuses on superficial cues the most as they seem to believe that to be a big indicator of what a prompt-based model bases its answers on. As for citations of the other two authors, Ryo Takahashi has 447 citations and Yusuke Oda has 1,171 citations. Yusuke Oda probably has the most citations given that he is a main researcher at his university and an editor of the Japanese Association for Natural Language Processing. Which means he is likely involved to some degree in a large amount of research and studies.

**References:**


- Pride Kavumba, Ryo Takahashi, and Yusuke Oda. 2022. Are Prompt-based Models Clueless?. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2333–2352, Dublin, Ireland. Association for Computational Linguistics.