Eric Zhang
Title of paper: GPTEval: A Survey on Assessments of ChatGPT and GPT-4

This paper seeks to go over previous assessments of ChatGPT and GPT-4 relating to their ethics, reasoning and language abilities, and scientific knowledge. It also has the goal of reviewing the current methods of evaluating these models to provide recommendations for research in the future on evaluating these large language models.The researchers chose GPT-4 and ChatGPT due to many various evaluations judging these models to be state-of-the-art. From the study, it was found that these models demonstrated a strong performance in language generation and understanding as they were proficient in interacting with users through dialogue and can handle various NLP problems and explain their approach. However, it was found that these models are not fully comprehensive when compared to models with domain-specific knowledge. For general knowledge on science, ChatGPT was able to provide open responses, but it did struggle with questions that require multi-step reasoning. For evaluations methods, it was found that current methods rely on benchmark datasets and prompt engineering which may not be reliable due to different prompts resulting in different evaluations.

One area of weakness is the fact that for the ethics section at the end, there were mentions of various ethical problems relating to the generation of AI content for training and other ethical concerns concerning the training of AI models. However, the paper does not propose any potential solutions to these issues. In terms of evaluation, there is some weakness in just suggesting transparency in prompt design and comparison for comparing LLMs as it is difficult to completely remove bias from prompts when there are so many ways to word them in ways that could alter their performance. One strength that this paper has is how it goes in depth into the performance of ChatGPT through pulling from many sources which provides information on how it performs compared to other models and its performance on various types of tests such as writing jokes and generating summaries. In other words, the paper does a good job overall in comparing results from multiple sources to consider the different tests and kinds of evaluations that may influence the result of the studies analyzed.

One improvement that could be made is suggesting some ways of dealing with the ethical problems of training AI. For instance, one area that could be expanded on are ways to address the human-biased feedback that may mislead RLHF. Another area of improvement is suggesting a more standardized way of ensuring that there is no bias in prompt engineering on top of advocating for transparency.