ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope

Summary on evaluation of LLM

A recent study [1] evaluated the performance of ChatGPT, a large language model, on the United States Medical Licensing Exam (USMLE) and found that ChatGPT performed at or near the passing threshold for all three exams without any specialized training or reinforcement. This suggests the potential for large language models to assist with medical education and clinical decision making. Another paper [4] proposed a framework for quantitatively evaluating interactive language models (LLMs) like ChatGPT using publicly available datasets and found that ChatGPT outperformed LLMs with zero-shot learning on most tasks and even outperformed fine-tuned models on some tasks. The study also found that ChatGPT is better at understanding non-Latin script languages than generating them and suffers from hallucination problems like other LLMs, generating more extrinsic hallucinations from its parametric memory as it does not have access to an external knowledge base. Furthermore, the paper demonstrated the limitations of ChatGPT in reasoning categories under logical reasoning, non-textual reasoning, and commonsense reasoning, while performing well on tasks favoring reasoning capabilities. Nonetheless, ChatGPT's interactive feature enables human collaboration with the underlying LLM to improve its performance. The comprehensive evaluation highlights the importance of understanding the potential impacts and limitations of LLMs like ChatGPT in various domains.