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

Summary on evaluation of LLM

Language models have gained much attention with the introduction of ChatGPT, which has attracted interest from academics and industry due to its remarkable conversational competency, reasoning capabilities, and security features. In paper 1, researchers proposed a framework to evaluate ChatGPT's multitask, multilingual, and multimodal aspects. The authors found that ChatGPT is better at understanding non-Latin-script languages than generating them, with an accuracy of 63.41% on average in 10 different reasoning categories, making it an unreliable reasoner. ChatGPT suffers from hallucination problems and generates more extrinsic hallucinations from its parametric memory as it does not have access to an external knowledge base. In paper 2, researchers collected data to compare ChatGPT's responses and human experts in various areas. They conducted comprehensive evaluations and linguistic analyses of ChatGPT-generated content compared to that of humans. Afterward, the authors conducted extensive experiments on how to effectively detect whether a certain text is generated by ChatGPT or humans. In paper 4, researchers evaluated ChatGPT's performance on the United States Medical Licensing Exam, which suggests that large language models may have the potential to assist with medical education and clinical decision-making. Finally, in paper 5 researchers empirically analyzed the zero-shot learning ability of ChatGPT and found that it performs well on many tasks favoring reasoning capabilities, but still faces challenges when solving specific tasks such as sequence tagging. The papers highlight the key challenges, biases, and limitations of ChatGPT, providing directions for future research.