

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

Summary on evaluation of LLM

In recent years, large language models (LLMs) have been widely adopted in various fields such as education and medicine. For instance, ChatGPT, a widely popular AI tool, has shown promising results in its capability to perform complex tasks such as answering medical licensing exams without specialized training or reinforcement [1]. However, assessing the accuracy and reliability of ChatGPT and other LLMs remains a challenge. Fortunately, a number of frameworks have been proposed to evaluate the effectiveness and faithfulness of LLMs, such as FEQA, an automatic question-answering based metric for evaluating faithfulness in abstractive summarization [2]. Additionally, studies have found that ChatGPT performs well in many tasks favoring reasoning capabilities but still faces challenges in solving specific tasks such as sequence tagging [7]. Nevertheless, there are concerns over the limitations, bias, and ethical issues in using LLMs like ChatGPT in education and healthcare. Thus, more research is needed, keeping in mind the possible risks and limitations, to ensure that the use of LLMs in various fields is safe, ethical, and aligns with societal values and goals in the future.