Midterm

2024-03-23

Midterm - Using Red-Teaming LLMs

Within the past two years, we have seen the rise of complex large language models such as ChatGPT, Gemini, and Copilot. According to Reuters, ChatGPT reached 100 million active users only two months after launch, becoming the fastest growing software application ever. Individuals use LLMs to automate simple tasks, solve questions, provide feedback on works, and much more. Companies, throughout different industries including tech, retail and manufacturing, have also started implementing LLMs into their products. Suffice to say, LLMs and AI have taken over the world. With the rise of LLMs also comes the increasing risk of LLMs generating negative and harmful results. In order to know what prompts this response and prevent it, LLM researchers conduct a technique called red teaming where human testers deliberately attempt to trigger harmful outputs. However, using human testers as a red team is incredibly costly and time-consuming. In response, recent developments have attempted to use an outside LLM as a red team which is trained through reinforcement language. However, these red team LLMs are incredibly limited in their coverage of test cases and only produce a small number of effective test cases. Ultimately, this leads us to the following research paper. Researchers at MIT decided to continue using separate LLMs as red teams, but they trained them through a method called curiosity-driven reinforcement language. To give some background, in traditional reinforcement language (RL), an agent (in this case, the red team LLM) learns to make decisions through trial and error. The agent is either rewarded or penalized for each decision until it finally reaches its goal. Usually, the goal of traditional reinforcement language is a minimized or maximized output - in this case, the goal for the red team LLM is to maximize toxic responses. In curiosity-driven reinforcement learning, the agent is rewarded for pursuing curiosity and exploring its environment. In this paper, researchers trained the red team LLM to explore a wide variety of prompts with the hope of expanding coverage of test cases. After deploying the curiosity-driven red teaming (CRT), the researchers found that it provided an increase of diversity of responses. This stemmed from exploring prompts that did not seem immediately obvious or that were different from other toxic and inappropriate prompts. In order to calculate the quality of the output, the researchers constructed a toxicity metric where they collected the percentage of toxic responses from each prompt given to the CRT. They used an AI-powered hate speech classifier to classify responses as “safe” and “toxic.” Along with having an increase in diversity, it also had high levels of toxicity. In tests with traditional RL models and human red teams, it outperformed both groups in diversity and toxicity levels. While having clear benefits, optimizing red teaming LLMs raises some ethical and moral concerns. At the center of this argument: does the ends justify the means? To prevent toxic outputs in LLMs, we must expose it to toxic inputs and in this case, encourage it. Even though it being exposed to prompts that may lead to toxic outputs, it cannot be guaranteed that these prompts won’t be met.