JUL 01
Grace Harmon

OpenAI unveils CriticGPT to combat coding errors in ChatGPT

Article

The news: **OpenAI** is addressing errors in **ChatGPT's** coding abilities with a new generative AI (genAI) model to check its work.

- The new model, called **CriticGPT**, writes critiques of ChatGPT's code output to help AI trainers identify hallucinations and bugs more easily and reliably.

- OpenAI said that when trainers use CriticGPT to review ChatGPT code **they outperform trainers who aren't using CriticGPT 60% of the time**.
- The tool is being used internally to improve ChatGPT's functions and isn't available to the public.

Fact-checking: OpenAI will build CriticGPT into an existing technique called reinforcement learning from human feedback (RLHF) to train and fine-tune ChatGPT.

- RLHF includes collecting comparisons from human AI trainers to rate different ChatGPT responses against each other and teach the language model how to better interact with people and formulate appropriate responses.
- OpenAI has acknowledged that as genAI models get smarter, it gets harder for human AI trainers to notice errors or determine the most appropriate output to a given user's request.

Public perception: OpenAI has had a few rocky months in the spotlight, including multiple lawsuits over copyright infringement from media companies, negative press about dissolving its safety Superalignment team, and former employees coming forward with concerns about OpenAI's "reckless" pursuit of rapid growth.

Creating CriticGPT shows that OpenAI knows its product is imperfect and needs a tangible resource to improve it. But there are limitations: The company conceded that CriticGPT was only trained on very short ChatGPT answers and that it will need to develop new methods for longer and more complex tasks.

Key takeaway: The public opinion of genAI platforms like ChatGPT is clouded by concerns about accuracy, privacy, and ethics. CriticGPT is a step in the right direction for earning trust and improving its models' basic functions.