MediaWell

News Item

Language models might be able to self-correct biases—if you ask them | MIT Technology Review

By Niall Firth
March 22, 2023

> Large language models are infamous for spewing toxic biases, thanks to the reams of awful human-produced content they get trained on.
>
> But if the models are large enough, and humans have helped train them, then they may be able to self-correct for some of these biases. Remarkably, all we have to do is ask.

> That's the finding of an experiment out of AI lab Anthropic, described in a non-peer-reviewed paper, which analyzed large language models that had been trained using reinforcement learning from human feedback (RLHF), a technique that gets humans to steer the AI model toward more desirable answers.

Source: Language models might be able to self-correct biases—if you ask them | MIT Technology Review