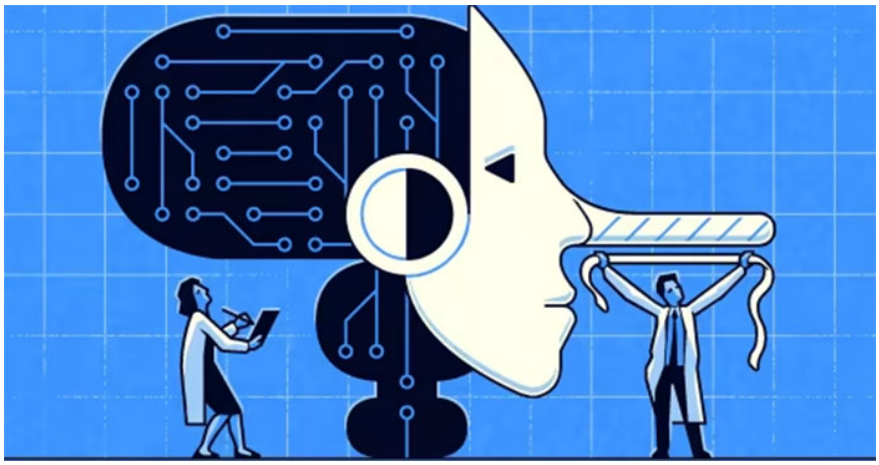
Nvidia develops 'AI police' that prevents ChatGPT from fabricating answers

Nvidia has announced a new software called NeMo Guardrails that has the ability to force super AIs like OpenAI's GPT or Google's LaMDA... not to give rambling answers, compose their own, and avoid toxic topics.

Nvidia has announced a new software called NeMo Guardrails that has the ability to force super AIs like OpenAI's GPT or Google's LaMDA. not to give rambling answers, compose their own, and avoid toxic topics. This software is considered one of the important steps to solve the problem of super AI having "hallucinations".



Super AIs like ChatGPT and Bard are trained to answer like humans, but they have a bad tendency to create answers that are somewhat silly, even dangerous.

NeMo Guardrails is the software layer that sits between users and super AIs. By adding multiple layers of filtering the AI's deemed malicious results, preventing the AI ??from talking about topics beyond its understanding but only talking about the topics the user is talking about or referring to. This reduces the likelihood of the AI ??providing made-up answers, removes malicious content, and limits the execution of harmful commands on the computer.

For example, NeMo Guardrails will limit a customer service chatbot designed to talk about company products to questions related to internal personnel, company security information, and products. of competitors even if customers ask.

Nvidia's "AI Police" tests the super AI's made-up answers using another large language model. Nemo Guardrails will not display content to users if the chatbot does not provide appropriate answers.

NeMo Guardrails is provided by Nivdia as open source through its services and can be used in commercial applications.