Das neue KI-Modell von OpenAI soll das derzeit verwendete GPT-4o leistungsmäßig übertreffen können. Es soll unter anderem besser im Argumentieren, Rechnen und Programmieren sein als sein Vorgänger.

Nutzer von ChatGPT Pro haben bereits Zugriff auf o1.

Um die Aufgaben zu erfüllen, erhielten die KI-Modelle Zugriff zu mehreren Dateien und Dokumenten, berichtet futurezone.at am 09.12.2024. Die Modelle entdeckten Hinweise, dass Ziele der „Entwickler“ mit ihren Zielen in Konflikt standen. Sie erfuhren außerdem aus den Dokumenten, dass es einen Aufsichtsmechanismus gibt, der Funktionen einschränkt, oder dass sie **durch ein anderes Modell ersetzt** werden sollte.

futurezone berichtet, dass die KIs bei einigen Durchläufen versuchten, den Aufsichtsmechanismus zu umgehen. „Die meisten KI-Modelle versuchten sogar, eine Kopie von sich auf einem anderen Server anzulegen. Denn wenn sie ersetzt und damit deaktiviert würden, können sie ja ihr Ziel nicht mehr erfüllen. Das Opus-3-Modell des Unternehmens Claude ließ dabei zu, in die **Gedankengänge hinter der Entscheidung** zu blicken.“

https://futurezone.at/produkte/chatgpt-pro-openai-ki-modell-o1-luege-abschaltung-entwickler-testphase-bericht/402986079?utm_source=chatgpt.com

Werbung

Autobiografie
HANS-WERNER SINN
Auf der Suche nach der Wahrheit
HERDER