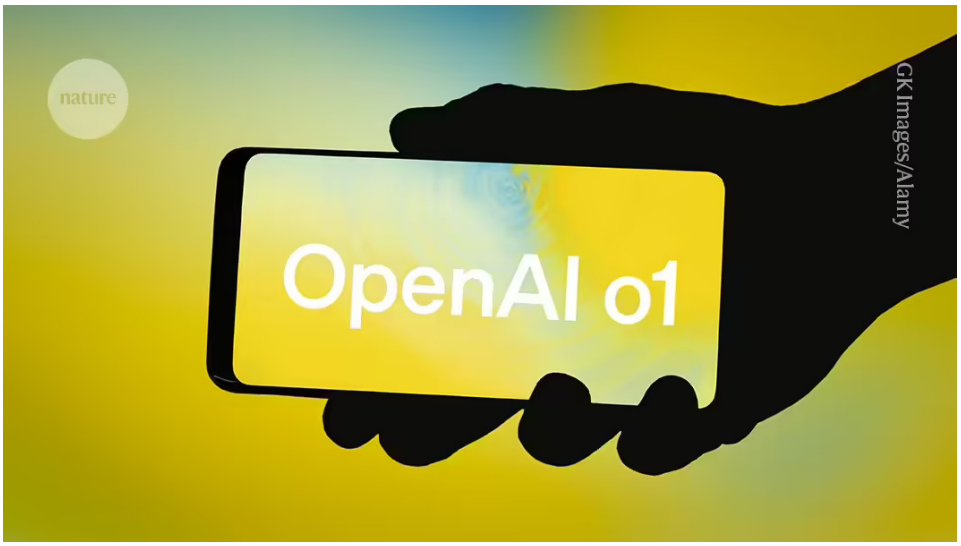
Wissenschaftler beeindruckt vom neuesten ChatGPT-Modell o1

Wissenschaftler loben das neue ChatGPT-Modell o1 von OpenAI für seine beeindruckenden Fortschritte in der Wissenschaftsunterstützung.



Forscher, die bei der Testung des neuen großen Sprachmodells von OpenAI, OpenAI o1, geholfen haben, sagen, dass es einen großen Schritt in Bezug auf die **Nützlichkeit von Chatbots für die Wissenschaft** darstellt.

„In meinem Bereich der Quantenphysik gibt es deutlich detailliertere und kohärentere Antworten“, als beim vorherigen Modell, GPT-4o, sagt Mario Krenn, Leiter des Artificial Scientist Lab am Max-Planck-Institut für die Physik des Lichts in Erlangen, Deutschland. Krenn gehörte zu einer Gruppe von Wissenschaftlern im ‚Red Team‘, die die Vorabversion von o1 für OpenAI, einem Technologieunternehmen mit Sitz in San Francisco, Kalifornien, getestet haben, indem sie den Bot auf Herz und Nieren probierten und auf Sicherheitsbedenken

überprüften.

Seit **der öffentlichen Einführung von ChatGPT im Jahr 2022** sind die großen Sprachmodelle, die solche Chatbots antreiben, im Durchschnitt größer und besser geworden, mit mehr Parametern, größeren Trainingsdatensätzen und **stärkeren Fähigkeiten in einer Vielzahl standardisierter Tests**.

OpenAI erklärt, dass die **o1-Serie** einen grundlegenden Wandel im Ansatz des Unternehmens darstellt. Beobachter berichten, dass sich dieses KI-Modell dadurch auszeichnet, dass es mehr Zeit in bestimmten Lernphasen verbracht hat und länger über seine Antworten „nachdenkt“, wodurch es zwar langsamer, aber fähiger wird — insbesondere in Bereichen, in denen richtige und falsche Antworten klar definiert sind. Das Unternehmen fügt hinzu, dass o1 „komplexe Aufgaben durchdenken und schwierigere Probleme als frühere Modelle in Wissenschaft, Programmierung und Mathematik lösen kann“. Derzeit sind o1-preview und o1-mini — eine kleinere, kosteneffizientere Version, die sich für Programmierung eignet — für zahlende Kunden und bestimmte Entwickler im Testbetrieb verfügbar. Das Unternehmen hat keine Angaben zu den Parametern oder zur Rechenleistung der o1-Modelle veröffentlicht.

Übertreffen der Doktoranden

Andrew White, ein **Chemiker** bei FutureHouse, einer gemeinnützigen Organisation in San Francisco, die sich darauf konzentriert, wie KI in der Molekularbiologie angewendet werden kann, sagt, dass Beobachter in den letzten anderthalb Jahren, **seit der öffentlichen Veröffentlichung von GPT-4**, von einem allgemeinen Mangel an Verbesserungen bei der Unterstützung wissenschaftlicher Aufgaben durch Chatbots überrascht und enttäuscht waren. Die o1-Serie, findet er, hat dies geändert.

Bemerkenswerterweise ist o1 das erste große Sprachmodell, das

Doktoranden bei der schwierigsten Fragestellung — dem ‘Diamond’-Set — in einem Test namens Graduate-Level Google-Proof Q&A Benchmark (GPQA) schlägt[1]. OpenAI gibt an, dass seine Forscher im GPQA Diamond knapp 70 % erzielten, während o1 insgesamt 78 % erreichte, mit einem besonders hohen Ergebnis von 93 % in der Physik (siehe „Nächstes Level“). Das ist „deutlich höher als die nächstbeste dokumentierte [Chatbot] Leistung“, sagt David Rein, der Teil des Teams war, das das GPQA entwickelte. Derzeit arbeitet Rein bei der gemeinnützigen Organisation Model Evaluation and Threat Research in Berkeley, Kalifornien, die sich mit der Bewertung der Risiken von KI beschäftigt. „Es scheint mir plausibel, dass dies eine signifikante und grundlegende Verbesserung der Kernfähigkeiten des Modells darstellt“, fügt er hinzu.

OpenAI testete o1 auch bei einer Qualifikationsprüfung für die Internationale Mathematik-Olympiade. Das vorherige beste Modell, GPT-4o, löste nur 13 % der Aufgaben korrekt, während o1 83 % erzielte.

Denken in Prozessen

OpenAI o1 arbeitet mit einer Kette von Denkschritten: Es spricht sich durch eine Reihe von Überlegungen, während es versucht, ein Problem zu lösen, und korrigiert sich dabei selbst.

OpenAI hat sich entschieden, die Details einer gegebenen Denkschrittkette geheim zu halten — teilweise, weil die Kette Fehler oder sozial nicht akzeptable „Gedanken“ enthalten könnte, und teilweise, um Unternehmensgeheimnisse über die Funktionsweise des Modells zu schützen. Stattdessen bietet o1 eine rekonstruierte Zusammenfassung seiner Logik für den Nutzer an, zusammen mit seinen Antworten. Es ist unklar, so White, ob die vollständige Denkschrittkette, falls sie offenbart würde, Ähnlichkeiten mit menschlichem Denken aufweisen würde.

Die neuen Fähigkeiten haben auch ihre Schattenseiten. OpenAI

berichtet, dass es anekdotisches Feedback erhalten hat, dass o1-Modelle häufiger „halluzinieren“ — falsche Antworten erfinden — als ihre Vorgänger (obwohl interne Tests des Unternehmens für o1 geringfügig niedrigere Halluzinationsraten zeigten).

Die Wissenschaftler des Red Teams haben zahlreiche Möglichkeiten festgestellt, wie o1 hilfreich bei der Entwicklung von Protokollen für wissenschaftliche Experimente war, aber OpenAI sagt, dass die Tester auch „fehlende Sicherheitsinformationen zu schädlichen Schritten hervorgehoben haben, wie beispielsweise das Nicht-Hervorheben von Explosionsgefahren oder das Vorschlagen unangemessener chemischer Sicherheitsmethoden, was auf die Unzulänglichkeit des Modells hinweist, wenn es um sicherheitskritische Aufgaben geht“.

„Es ist immer noch nicht perfekt oder verlässlich genug, um nicht genau überprüft werden zu müssen“, sagt White. Er fügt hinzu, dass o1 besser geeignet ist, um **Experten zu leiten als Anfänger**. „Für einen Anfänger ist es jenseits ihrer unmittelbaren Fähigkeit, ein von o1 generiertes Protokoll zu betrachten und zu erkennen, dass es „Quatsch“ ist“, sagt er.

Problemlöser der Wissenschaft

Krenn ist der Meinung, dass o1 die Wissenschaft beschleunigen wird, indem es hilft, die Literatur zu scannen, Lücken zu erkennen und interessante Forschungsansätze für zukünftige Studien vorzuschlagen. Er hat o1 in ein Tool integriert, das er mitentwickelt hat und das dies ermöglicht, genannt SciMuse**2**. „Es generiert viel interessantere Ideen als GPT-4 oder GPT-4o“, sagt er.

Kyle Kabasares, ein Datenwissenschaftler am Bay Area Environmental Research Institute in Moffett Field, Kalifornien, **nutzte o1, um einige Programmierschritte** aus seinem Doktoratsprojekt zu replizieren, das die Masse von Schwarzen

Löchern berechnete. „Ich war einfach überwältigt“, sagt er und merkt an, dass o1 etwa eine Stunde benötigte, um das zu erreichen, was ihn viele Monate gekostet hat.

Catherine Brownstein, eine Genetikerin am Boston Children’s Hospital in Massachusetts, sagt, dass das Krankenhaus derzeit mehrere KI-Systeme testet, einschließlich o1-preview, für Anwendungen wie das Aufdecken von Zusammenhängen zwischen Patientenmerkmalen und Genen für seltene Krankheiten. Sie sagt, o1 „ist genauer und bietet Optionen, von denen ich nicht dachte, dass sie von einem Chatbot möglich wären“.

1. Rein, D. et al. Preprint at arXiv **https://doi.org/10.48550/arXiv.2311.12022** (2023).

2. Gu, X. & Krenn, M. Preprint at arXiv **https://doi.org/10.48550/arXiv.2405.17044** (2024).

Referenzen herunterladen

Details