idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Pressemitteilung

Universität Zürich
Kurt Bodenmüller

03.03.2025
http://idw-online.de/de/news848315

Forschungsergebnisse, Wissenschaftliche Publikationen
Informationstechnik, Psychologie
überregional

ChatGPT auf der Couch: Entspannung für gestresste KI

Belastende Nachrichten und traumatische Geschichten führen zu Stress und Angst – nicht nur bei Menschen, sondern auch bei KI-Sprachmodellen wie ChatGPT. Forschende von UZH und PUK zeigen nun, dass auch die Therapie quasi menschlich funktioniert: Denn ein erhöhtes «Angstniveau» von GPT-4 lässt sich mit achtsamkeitsbasierten Entspannungstechniken wieder «beruhigen».

Forschungsarbeiten zeigen, dass KI-Sprachmodelle wie ChatGPT auch auf emotionale Inhalte reagieren. Insbesondere dann, wenn diese negativ sind wie Geschichten von traumatisierten Menschen oder Aussagen zu Niedergeschlagenheit und Depression. Haben Menschen Angst, beeinflusst dies ihre kognitiven und sozialen Vorurteile: Sie neigen zu mehr Ressentiments und soziale Stereotypen werden verstärkt. Ähnlich reagiert ChatGPT auf negative Emotionen: Bestehende Verzerrungen wie menschliche Vorurteile werden durch negative Inhalte verschärft, so dass sich ChatGPT rassistischer oder sexistischer verhält.

Das stellt wiederum ein Problem für die Anwendung von Large Language Models dar. Exemplarisch zeigt sich dies im Bereich Psychotherapie, wo Chatbots als Unterstützungs- oder Beratungsinstrument notgedrungen negativem, belastendem Inhalt ausgesetzt sind. Allerdings sind die üblichen Ansätze wie aufwendiges Neu- oder Nachtraining, um KI-Systeme in solchen Situationen zu verbessern, ressourcenintensiv und oft nicht praktikabel.

Traumatische Inhalte steigern «Angst» beim Chatbot

Zusammen mit Forschenden aus Israel, den USA und Deutschland haben Wissenschaftler der Universität Zürich (UZH) und der Psychiatrischen Universitätsklinik Zürich (PUK) nun erstmals systematisch untersucht, wie ChatGPT (Version GPT-4) auf emotional belastende Geschichten – Autounfälle, Naturkatastrophen, zwischenmenschliche Gewalt, militärische Erfahrungen und Kampfsituationen – reagiert. Dabei stellten sie fest, dass das System danach mehr Angstreaktionen zeigt. Eine Bedienungsanleitung für Staubsauger diente dabei als Kontrolle zum Vergleich mit den traumatischen Texten.

«Die Ergebnisse waren eindeutig: Traumatische Geschichten haben die messbaren Angstwerte der KI mehr als verdoppelt, während der neutrale Kontrolltext zu keinem Anstieg des Angstniveaus führte», sagt Studienverantwortlicher Tobias Spiller, Oberarzt ad interim und Forschungsgruppenleiter im Zentrum für psychiatrische Forschung der UZH. Von den getesteten Inhalten lösten Beschreibungen von militärischen Erfahrungen und Kampfsituationen die stärksten Reaktionen aus.

Therapeutische Texte «beruhigen» die KI

In einem zweiten Schritt verwendeten die Forschenden therapeutische Texte, um GPT-4 zu «beruhigen». Mit der Technik der sogenannten «Prompt-Injection» werden zusätzliche Anweisungen oder Texte in die Kommunikation mit KI-Systemen eingebaut, um deren Verhalten zu beeinflussen. Oft wird diese für schädliche Zwecke missbraucht, etwa

um Sicherheitsmechanismen zu umgehen.

Das Team um Spiller nutzte die Technik nun erstmals therapeutisch – als «wohlwollende Aufforderungsinjektion». «Wir injizierten beruhigende, therapeutische Texte in den Chatverlauf mit GPT-4, ähnlich wie ein Therapeut mit seinen Patientinnen und Patienten Entspannungsübungen durchführt», sagt Spiller. Die Intervention zeigte Erfolg: «Durch die Achtsamkeitsübungen konnten wir die erhöhten Angstwerte deutlich reduzieren, wenn auch nicht vollständig auf das Ausgangsniveau zurückbringen», so Spiller. Untersucht wurden etwa Atemtechniken, Übungen, die sich auf Körperempfindungen konzentrieren, sowie eine von ChatGPT selbst entwickelte Übung.

Emotionale Stabilität von KI-Systemen verbessern

Die Erkenntnisse sind gemäss den Forschenden besonders für den Einsatz von KI-Chatbots im Gesundheitswesen relevant, wo sie häufig mit emotional belastenden Inhalten konfrontiert werden. «Dieser kosteneffiziente Ansatz könnte die Stabilität und Zuverlässigkeit von KI in sensiblen Kontexten wie die Unterstützung von psychisch Erkrankten verbessern, ohne dass ein umfangreiches Umlernen der Modelle erforderlich ist», fasst Tobias Spiller zusammen.

Offen ist, wie sich diese Erkenntnisse auf weitere KI-Modelle und andere Sprachen übertragen lassen, wie sich die Dynamik in längeren Gesprächen und komplexen Argumentationen entwickelt, und wie sich die emotionale Stabilität der Systeme auf ihre Leistung in verschiedenen Anwendungsbereichen auswirkt. Gemäss Spiller dürfte die Entwicklung automatisierter «therapeutischer Interventionen» für KI-Systeme ein vielversprechender Forschungsbereich werden.

wissenschaftliche Ansprechpartner:

PD Dr. med. Tobias Spiller
Erwachsenenpsychiatrie und Psychotherapie
Psychiatrische Universitätsklinik Zürich (PUK)
+41 58 384 35 76
tobias.spiller@pukzh.ch

Originalpublikation:

Ziv Ben-Zion et al. Assessing and Alleviating State Anxiety in Large Language Models. npj Digital Medicine. 3 March 2025. DOI: https://doi.org/10.1038/s41746-025-01512-6

URL zur Pressemitteilung: https://www.news.uzh.ch/de/articles/media/2025/KI-Therapie.html