MedWiss.Online

Eignen sich KI-Chatbots fürs Krankenhaus?

Diagnose-Fähigkeiten von Large Language Models getestet

Large Language Models bestehen medizinische Examen mit Bravour. Sie für Diagnosen heranzuziehen, wäre derzeit aber grob fahrlässig: Medizin-Chatbots treffen vorschnelle Diagnosen, halten sich nicht an Richtlinien und würden das Leben von Patientinnen und Patienten gefährden. Zu diesem Schluss kommt ein Team der TUM, das erstmals systematisch untersucht hat, ob diese Form der Künstlichen Intelligenz (KI) für den Klinikalltag geeignet wäre. Die Forschenden sehen dennoch Potenzial in der Technologie. Sie haben ein Verfahren veröffentlicht, mit dem sich die Zuverlässigkeit zukünftiger Medizin-Chatbots testen lässt.

Large Language Models sind Computerprogramme, die mit riesigen Mengen Text trainiert wurden. Speziell trainierte Varianten der Technologie, die auch hinter ChatGPT steckt, lösen mittlerweile sogar Abschlussexamen aus dem Medizinstudium nahezu fehlerfrei. Wäre eine solche KI auch in der Lage, die Aufgaben von Ärztinnen und Ärzten in einer Notaufnahme zu übernehmen? Könnte sie anhand der Beschwerden die passenden Tests anordnen, die richtige Diagnose stellen und einen Behandlungsplan entwerfen?

Im Fachmagazin „Nature Medicine“ hat sich ein interdisziplinäres Team um Daniel Rückert, Professor für Artificial Intelligence in Healthcare and Medicine an der TUM, dieser Frage gewidmet. Ärztinnen und Ärzte haben gemeinsam mit KI-Fachleuten erstmals systematisch untersucht, wie erfolgreich verschiedene Varianten des Open-Source-Large-Language-Models Llama 2 bei der Diagnose sind.

Weg von Notaufnahme zur Behandlung nachgespielt

Um die Fähigkeiten der komplexen Algorithmen zu testen, nutzten die Forschenden anonymisierte Daten von Patientinnen und Patienten aus einer Klinik in den USA. Aus einem größeren Datensatz wählten sie 2.400 Fälle aus. Alle Betroffenen waren mit Bauchschmerzen in die Notaufnahme gekommen. Die Fallbeschreibung endete jeweils mit einer von vier Diagnosen und einem Behandlungsplan. Zu den Fällen waren alle Daten verfügbar, die für die Diagnose erfasst wurden – von der Krankengeschichte über die Blutwerte bis hin zu den Bildgebungsdaten. „Wir haben die Daten so aufbereitet, dass die Algorithmen die realen Abläufe und Entscheidungsprozesse im Krankenhaus nachspielen konnten“, erläutert Friederike Jungmann, Assistenzärztin in der Radiologie des Klinikums rechts der Isar der TUM und gemeinsam mit dem Informatiker Paul Hager Erstautorin der Studie. „Das Programm hat immer nur die Informationen, die auch die realen Ärztinnen und Ärzte hatten. Ob es beispielsweise ein Blutbild in Auftrag gibt, muss es selbst entscheiden und dann mit dieser Information die nächste Entscheidung treffen, bis es schließlich eine Diagnose und einen Behandlungsplan erstellt.“

Das Team stellte fest, dass keines der Large Language Models durchgängig alle notwendigen Untersuchungen einforderte. Tatsächlich wurden die Diagnosen der Programme sogar weniger zutreffend, je mehr Informationen sie zu dem Fall hatten. Behandlungsrichtlinien befolgten sie oftmals nicht. Als Konsequenz ordnete die KI beispielsweise Untersuchungen an, die für echte Patientinnen und Patienten schwere gesundheitliche Folgen nach sich gezogen hätten.

Direkter Vergleich mit Ärztinnen und Ärzten

In einem zweiten Teil der Studie wurden KI-Diagnosen zu einer Teilmenge aus dem Datensatz mit Diagnosen von vier Ärztinnen und Ärzten verglichen. Während diese bei 89 Prozent der Diagnosen richtig lagen, kam das beste Large Language Model auf gerade einmal 73 Prozent. Jedes Modell erkannte manche Erkrankungen besser als andere. In einem Extremfall diagnostizierte ein Modell Gallenblasenentzündungen nur in 13 Prozent der Fälle korrekt.

Ein weiteres Problem, das die Programme für den Einsatz im Alltag disqualifiziert, ist ein Mangel an Robustheit: Welche Diagnose ein Large Language Modell stellte, hing unter anderem davon ab, in welcher Reihenfolge es die Informationen erhielt. Auch linguistische Feinheiten beeinflussten das Ergebnis – beispielsweise ob das Programm um eine „Main Diagnosis“, eine „Primary Diagnosis“ oder eine „Final Diagnosis“ gebeten wurde. Im Klinikalltag sind die Begriffe in der Regel austauschbar.

ChatGPT nicht getestet

Das Team hat explizit nicht die kommerziellen Large Language Models von OpenAI (ChatGPT) und Google getestet. Dafür gibt es im Wesentlichen zwei Gründe. Zum einen untersagt der Anbieter der Krankenhausdaten aus Datenschutzgründen, die Daten mit diesen Modellen zur verarbeiten. Zum anderen raten Expertinnen und Experten nachdrücklich, für Anwendungen im Gesundheitssektor ausschließlich Open-Source-Software zu verwenden.

„Nur mit Open-Source-Software haben Krankenhäuser die Informationen und das nötige Maß an Kontrolle, um die Sicherheit der Patientinnen und Patienten zu gewährleisten. Wenn es darum geht, Large Language Models zu bewerten, müssen wir wissen, mit welche Daten sie trainiert wurden. Sonst könnte es sein, dass wir für die Bewertung genau die Fragen und Antworten verwenden, mit denen sie trainiert wurden. Da Unternehmen die Trainingsdaten streng unter Verschluss halten, würde eine faire Bewertung erschwert”, sagt Paul Hager. “Es ist auch gefährlich, wichtige medizinische Infrastrukturen von externen Dienstleistern abhängig zu machen, die ihre Modelle nach Belieben aktualisieren und ändern können. Im Extremfall könnte ein Dienst, den Hunderte von Kliniken nutzen, eingestellt werden, weil er nicht mehr rentabel ist.”

Schnelle Fortschritte

Die Entwicklung in dieser Technologie verläuft sehr schnell. „Es ist gut möglich, dass in absehbarer Zeit ein Large Language Model besser dafür geeignet ist, aus Krankengeschichte und Testergebnissen auf eine Diagnose zu kommen“, sagt Prof. Daniel Rückert. „Wir haben deshalb unsere Testumgebung für alle Forschungsgruppen freigegeben, die Large Language Models für den Klinikkontext testen wollen.“ Rückert sieht Potenzial in der Technologie: „Large Language Models könnten in Zukunft wichtige Werkzeuge für Ärztinnen und Ärzte werden, mit denen sich beispielsweise ein Fall diskutieren lässt. Wir müssen uns aber immer der Grenzen und Eigenheiten dieser Technologie bewusst sein und diese beim Erstellen von Anwendungen berücksichtigen“, sagt der Medizin-KI-Experte.

Publikationen

Hager, P., Jungmann, F., Holland, R. *et al.* “Evaluation and mitigation of the limitations of large language models in clinical decision-making”. *Nat Med* (2024). DOI: 10.1038/s41591-024-03097-1

Weitere Informationen und Links

Prof. Daniel Rückert ist einer der Direktoren des Munich Data Science Institute (MDSI) und Leiter

des Zentrums für Digitale Medizin und Gesundheit an der TUM.