MedWiss.Online

Zu vorsichtig für die Versorgung: Schwächen von ChatGPT bei Gesundheitsfragen

Laut einer Studie der TU Berlin neigen ChatGPT-Modelle zu übervorsichtigen Empfehlungen. Für die gezielte Steuerung von Patient*innen im Gesundheitssystem reicht das aktuell nicht aus

Künstliche Intelligenz (KI) wird zunehmend auch für gesundheitliche Fragen genutzt. Viele Menschen verwenden Tools wie ChatGPT, um Beschwerden einzuordnen und abzuschätzen, ob sie sofort medizinische Hilfe brauchen, ärztlichen Rat einholen sollten oder zunächst abwarten können. Mit speziell für den Gesundheitsbereich positionierten Versionen wie etwa ChatGPT Health in den USA entsteht dabei leicht der Eindruck besonderer fachlicher Eignung. Wie verlässlich Empfehlungen von ChatGPT tatsächlich sind, ist bislang jedoch nur begrenzt untersucht.

In einer neuen Studie aus dem Fachgebiet Arbeitswissenschaft der Technischen Universität Berlin haben Forschende deshalb analysiert, wie genau ChatGPT in verschiedenen Modellversionen gesundheitliche Beschwerden einordnet, wie sich die Leistung im Zeitverlauf verändert hat und ob identische Eingaben konsistente Empfehlungen erzeugen. Das Ergebnis: Für die digitale Ersteinschätzung und eigenständige Patientensteuerung ist ChatGPT derzeit nur eingeschränkt geeignet.

22 Modellversionen, 45 reale Fälle, 9.900 Bewertungen

„Der Hauptunterschied zu unseren früheren Studien ist die längsschnittliche Analyse. Bisher wurden nur ein oder zwei Modelle untersucht. Nun haben wir alle Modelle, die über die Zeit verfügbar waren, getestet und analysiert, wie sie sich tatsächlich verändert haben“, sagt Studienleiter Dr. Marvin Kopka. „Das war uns auch deshalb wichtig, weil es immer wieder Meldungen gibt, nach denen neue Modelle in ärztlichen Zulassungsprüfungen oder Wissenstests nahezu perfekte Ergebnisse erreichen. Daraus wird dann schnell geschlossen, dass sie auch für Patient*innen verlässliche medizinische Empfehlungen geben. Genau das stimmt aber laut unserer Studie nicht.“

Für die Studie „Evaluating the accuracy of ChatGPT model versions for giving care-seeking advice“, erschienen im Journal „Communications Medicine“, testete das Forschungsteam 22 ChatGPT-Modellversionen anhand echter Fälle von 45 Patient*innen. Darunter waren Krankheitsbilder wie „eine kurzfristige Überlastung von Sehnen/Bändern am Vortag“ oder auch „einfache Verdauungsprobleme/Durchfall seit einem Tag ohne weitere Beschwerden“. Jeder Fall wurde pro Modell zehnmal eingegeben. Insgesamt entstanden so 9.900 Einzelbewertungen. Die Modelle mussten jeweils entscheiden, ob ein Fall als Notfall, als Fall für ärztliche Abklärung oder als Fall für Selbstversorgung einzustufen ist.

Die Genauigkeit steigt kaum noch

Die Auswertung zeigt: Die Genauigkeit stieg mit den ersten Modellversionen zunächst deutlich an. Seit der dritten Modellgeneration (gpt-4) gab es jedoch nur noch geringfügige Verbesserungen. Das beste getestete Modell erreichte eine Treffergenauigkeit von 74 Prozent. Zwar empfahlen neuere Modelle häufiger überhaupt Selbstversorgung, insgesamt blieb die Leistung in diesem Bereich aber

begrenzt.

Besondere Schwächen bei harmlosen Beschwerden

Besonders gut waren die getesteten Modelle darin, behandlungsbedürftige Fälle zu erkennen. Die meisten Fehler traten dagegen bei Fällen auf, in denen Selbstversorgung ausreichend gewesen wäre: 70 Prozent aller Fehler entfielen auf diese Gruppe. Kein einziger der 13 Selbstversorgungsfälle wurde von allen Modellen in allen Durchläufen korrekt gelöst.

Lediglich einzelne Modelle, etwa o4, o3 oder GPT 5, empfahlen überhaupt jemals Selbstversorgung. Bei allen anderen getesteten Modellen wurde durchgängig zur ärztlichen Abklärung geraten. Das ist problematisch, weil ein erheblicher Teil der Beschwerden tatsächlich nicht gefährlich ist, von allein wieder weggeht oder selbst behandelt werden kann.
Die Studie zeigt damit ein strukturelles Muster: Fast alle Modelle neigen dazu, Beschwerden vorsichtshalber als behandlungsbedürftiger einzustufen, als es medizinisch erforderlich wäre.

Die Forschenden bezeichnen dieses Muster als konservatives Triagierungsverhalten. „Uns haben die Ergebnisse in dieser Klarheit selbst überrascht“, so Dr. Marvin Kopka. „Denn sie zeigen explizit, dass die für Patient*innen relevanten Fragen durch neuere Modelle nicht automatisch besser beantwortet werden. Bessere Test- oder Prüfungsergebnisse bedeuten eben nicht zwangsläufig einen höheren praktischen Nutzen in der Versorgung.“

Entscheidend ist der praktische Nutzen

„Entscheidend ist aus unserer Sicht nicht nur, ob ein Modell einzelne Fälle richtig einordnet, sondern welchen praktischen Nutzen die Empfehlungen im Alltag tatsächlich haben. Wenn ein System bei sehr vielen Beschwerden vorsorglich zur medizinischen Abklärung rät, wirkt das zunächst sicher für Nutzer*innen – es bietet aber faktisch keine echte Entscheidungshilfe mehr, wenn die Empfehlung fast immer gleich ausfällt“, so Dr. Marvin Kopka.

Gleiche Eingabe, nicht immer gleiche Empfehlung

Hinzu kommt ein weiteres Problem: Die Modelle antworten nicht durchgängig konsistent. Bei identischen Eingaben kam es je nach Modell zu teils deutlichen Schwankungen. Neuere Modelle hatten zwar seltener Fälle, die nie korrekt gelöst wurden, zugleich aber häufiger Fälle mit inkonsistenten Empfehlungen über mehrere Durchläufe hinweg. Besonders deutlich zeigte sich das bei GPT 5: Bei 42 Prozent aller Fälle waren die Empfehlungen bei mehrfacher Eingabe desselben Falls mal richtig und mal falsch – trotz exakt gleicher Eingabe.

Im Experiment zeigte sich zwar, dass sich die Genauigkeit verbessern lässt, wenn dieselbe Frage mehrfach gestellt und anschließend die niedrigste Dringlichkeitsstufe aus mehreren Antworten ausgewählt wird. Auf diese Weise stieg die Gesamtgenauigkeit im Mittel um vier Prozentpunkte, die Genauigkeit bei Selbstversorgungsfällen sogar um 14 Prozentpunkte. Die Forschenden betonen aber ausdrücklich, dass dies keine Empfehlung für Endnutzer*innen ist, weil dabei im schlimmsten Fall Notfälle übersehen werden könnten.

Relevanz für die Debatte um Primärversorgung

Die Ergebnisse sind auch gesundheitspolitisch relevant, so Kopka. In Deutschland wird intensiv über ein Primärversorgungssystem und über Formen digitaler Patientensteuerung diskutiert. Die TU-Studie legt nahe, dass allgemeine Sprachmodelle wie ChatGPT dafür derzeit kein geeignetes allein einsetzbares Instrument sind. Wenn ein System in der Praxis überwiegend zur ärztlichen Abklärung rät, entsteht kaum ein echter Steuerungseffekt – unnötige ärztliche Inanspruchnahme kann dann

sogar zunehmen.

Potenzial eher in qualitätsgesicherten Anwendungen

„Das Potenzial großer Sprachmodelle sehen wir deshalb derzeit weniger in einer Nutzung im Chatfenster der Hersteller als in einer sinnvollen Integration in qualitätsgesicherten Anwendungen, also in Symptom-Checker-Apps. Dort könnten sie helfen, Informationen verständlich aufzubereiten, Empfehlungen zu erläutern und Menschen besser durch bestehende Versorgungswege zu lotsen – vorausgesetzt, die medizinische Qualitätssicherung erfolgt im Hintergrund“, so Marvin Kopka.

Einschränkungen der Studie

Die Forschenden weisen zugleich daraufhin, dass der Fokus dieser Studie auf Bevölkerungsrepräsentativität lag. Da echte Notfälle im Alltag selten sind und dementsprechend auch seltener bei der Nutzung von ChatGPT auftreten, enthielt auch der Datensatz nur wenige Notfälle und untersuchte hauptsächlich Entscheidungen für oder gegen das Aufsuchen von ärztlicher Hilfe. Die Genauigkeit bei der Erkennung von echten Notfällen sollte in weiteren Studien untersucht werden.

Studie:

Kopka, M., He, L. & Feufel, M.A., Evaluating the accuracy of ChatGPT model versions for giving care-seeking advice. Commun Medicine
(2026). https://www.nature.com/articles/s43856-026-01466-0