MedWiss.Online

KI weist den Weg zur richtigen Therapie

Symptom-Checker aus dem Internet liegen häufig daneben. Kann KI es besser? Forschende aus dem Team von Altuna Akalin am MDC-BIMSB haben untersucht, inwieweit große Sprachmodelle Patient*innen beraten und Ärzt*innen unterstützen können. Ihre Studie ist in „npj Digital Medicine“ veröffentlicht.

Zuerst war da nur ein Stechen, ein merkwürdiges Gefühl in der Brust. Dann kam eine unerklärliche Müdigkeit hinzu. Sie sitzen auf dem Sofa und zögern. Sollen Sie ärztlichen Rat einholen? Oder warten Sie erst einmal ab? Und falls Sie eine Ärztin oder einen Arzt benötigen: Wäre eine kardiologische, eine internistische oder vielleicht doch eine neurologische Praxis die beste Anlaufstelle?

Wenn es Ihnen wie den meisten Menschen geht, suchen Sie als Erstes Rat im Internet. Zwar gibt es dort inzwischen zahlreiche Symptom-Checker, doch diese sind selten akkurat. Eine Studie aus dem Jahr 2022 hat ergeben, dass digitale Symptom-Checker die richtige Diagnose nur in 19 bis 38 Prozent der Fälle an erster Stelle nennen. Werden die ersten drei Vorschläge berücksichtigt, ist die passende Diagnose zwar häufiger dabei, aber auch nur in 33 bis 58 Prozent aller Anfragen.

In einer im Fachblatt „npj Digital Medicine“ veröffentlichten Studie hat ein Team um Farieda Gaber aus dem Labor von Dr. Altuna Akalin, dem Leiter der Technologieplattform „Bioinformatics and Omics Data Science“ am Berliner Institut für Medizinische Systembiologie des Max Delbrück Center (MDC-BIMSB), jetzt untersucht, ob große Sprachmodelle (Large Language Models, kurz LLMs) mehr leisten können. Genauer gesagt sind die Forschenden der Frage nachgegangen, wie gut LLMs Patient*innen und Ärzt*innen den Weg zur passenden Therapie weisen können.

„Studien zeigen, dass bis zu 30 Prozent der Besuche in Notaufnahmen nicht notwendig sind“, sagt Akalin, der korrespondierender Autor der Studie ist. „Wenn LLMs helfen könnten, diese Zahl zu reduzieren, würde das zur Entlastung der Gesundheitssysteme beitragen.“

Für seine Studie verglich das Team die Ergebnisse von vier Varianten von Claude, einem von der US-Firma Anthropic entwickelten Sprachmodell, mit 2.000 realen Fällen aus Notaufnahmen. Die Daten dazu entstammen der MIMIC-IV-ED-Datenbank, einer großen öffentlichen Sammlung anonymisierter Gesundheitsdaten aus dem Beth Israel Deaconess Medical Center in Boston.

Die Modelle sollten drei Dinge tun: die passende Fachärztin oder den passenden Facharzt vorschlagen, eine Diagnose stellen und die Dringlichkeit des Falles beurteilen. Diese Einschätzung wird auch als Triage bezeichnet. Es wurden zwei Szenarien durchgespielt. Das erste Szenario simulierte eine Situation, in der sich eine Patientin oder ein Patient zu Hause befand, sodass nur Symptome und demografische Daten vorlagen. Das zweite ahmte die Situation in einer ärztlichen Praxis nach, wodurch zusätzlich Vitalparameter wie Herzfrequenz und Blutdruck verfügbar waren.

Das passende Fachgebiet

Anders als in Deutschland, wo Patient*innen in der Regel eine Überweisung von ihrer Hausärztin oder ihrem Hausarzt benötigen, können Menschen in vielen anderen Ländern direkt eine fachärztliche Praxis aufsuchen. Es kann jedoch schwierig sein, herauszufinden, welche

Spezialist*innen für die jeweiligen Beschwerden am geeignetsten sind. Ist eine gastroenterologische Praxis die beste, wenn es um Bauchschmerzen geht? Oder wäre ein nephrologische die bessere Wahl?

Bei dieser Aufgabe erwiesen sich die LLMs als sehr zuverlässig. Wurden nur Symptome genannt, wählte das Modell Claude 3.5 Sonnet beispielsweise in etwa 87 Prozent der Fälle bei seinen ersten drei Vorschlägen ein geeignetes Fachgebiet aus. Die anderen Modelle schnitten ähnlich gut ab. Die Genauigkeit der LLMs verbesserte sich geringfügig, wenn sie zusätzliche Informationen zu den Vitalparametern der Patient*innen erhielten. Die Ärzt*innen, die die KI-Vorschläge überprüften, waren sich einig: Sie bewerteten 97 Prozent der Empfehlungen als genau oder zumindest als klinisch akzeptabel.

Die richtige Diagnose

Auch bei der Diagnose schnitten die Modelle gut ab. Die beste Version erkannte die richtige Erkrankung in mehr als 82 Prozent der Fälle. Die Genauigkeit stieg weiter, wenn die Vitalparameter vorlagen – insbesondere bei der RAG-Variante (Retrieval Augmented Generation), die bei ihrer Entscheidungsfindung auf eine Datenbank mit rund 30 Millionen PubMed-Abstracts zurückgreifen kann.

Wie gut die KI-Diagnosen mit dem menschlichen Urteil übereinstimmten, prüften die Forschenden auf zwei verschiedene Arten. In der einen Variante, bei der eine Vorhersage als richtig galt, wenn mindestens eine*r von zwei unabhängigen Ärzt*innen ihr zustimmte, war sich die KI in mehr als 95 Prozent der Fälle mit dem menschlichen Urteil einig. In der anderen, strengeren Variante, bei der beide Ärzt*innen dem KI-Urteil zustimmen mussten, betrug die Übereinstimmung immerhin gut 70 Prozent.

Die Triage bleibt knifflig

Bei der Beurteilung der Dringlichkeit eines Falls waren die Modelle weniger akkurat. Zwar verwechselte keines von ihnen einen lebensbedrohlichen Zustand mit einem harmlosen, aber mittelschwere Fälle schätzten sie oft falsch ein. Das ist wichtig, denn sowohl eine Übertriagierung – die Bevorzugung stabiler Patient*innen – als auch eine Untertriagierung – die verzögerte Behandlung schwerer Fälle – können den Betroffenen schaden. In der Notfallversorgung wird eine Abweichung von weniger als 5 Prozent angestrebt; dieses Ziel erreichte in der Studie keines der Modelle.

Auch hier schnitten die LLMs, die Zugang zu den Vitaldaten hatten, allerdings besser ab. Das deutet den Forschenden zufolge darauf hin, dass sich die Resultate, die per KI erzielt werden können, weiter verbessern lassen, wenn noch mehr in medizinischen Tests gewonnene Daten in die Modelle eingespeist werden.

KI kann Ärzt*innen nicht ersetzen, aber unterstützen

„Wir empfehlen natürlich nicht, Ärzt*innen durch KI-Tools zu ersetzen“, sagt Akalin. „Aber gut konzipierte, rigoros getestete LLMs könnten für Mediziner*innen eine hilfreiche Unterstützung sein, insbesondere für die noch weniger erfahrenen unter ihnen.“ Er und seine Kolleg*innen würden sich zudem wünschen, dass Patient*innen Zugriff auf bestimmte Arten von LLMs erhalten – insbesondere auf solche, die bei der Suche nach Fachärzt*innen helfen. Die Modelle könnten die weniger präzisen Symptom-Checker ersetzen und bei der Frage, ob und wo man sich behandeln lassen sollte, behilflich sein. Indem LLMs unnötige Arzt- und Krankenhausbesuche reduzieren, könnten sie zudem die Gesundheitssysteme entlasten, fügt Akalin hinzu.

Bevor solche Werkzeuge offiziell genutzt werden dürfen, müssen strenge regulatorische Standards gemäß dem EU-Gesetz über künstliche Intelligenz erfüllt sein. Dennoch warnen die Autor*innen vor einem unsicheren Einsatz, wenn öffentlich verfügbare KI-Tools informell im klinischen Umfeld eingesetzt werden. „Deshalb ist ein offenes, strenges Benchmarking so wichtig“, sagt die Erstautorin der Studie, Gaber. „Forschung wie diese hilft uns, sowohl die Möglichkeiten als auch die Grenzen von KI-gestützten medizinischen Entscheidungen zu verstehen.“

Akalin und sein Team planen, den Wert von LLMs sowohl für Patient*innen als auch für Ärzt*innen in realen Umgebungen, zum Beispiel in ärztlichen Praxen, mithilfe der in seinem Labor entwickelten Plattform 2ndOpin.io weiter zu testen. „Die nächste Frage lautet: Wenn wir ein solches Tool bauen, ist es dann wirklich nützlich?“, sagt er. LLMs, die die Patientenversorgung verbessern, sind ein Forschungsschwerpunkt von Akalin, der auch onconaut.ai entwickelt hat – ein KI-basiertes Online-Tool für Ärzt*innen und Patient*innen, das helfen kann, sich besser in personalisierten Krebstherapien zurechtzufinden. Krebspatient*innen können dort beispielsweise ihren Biomarker-Status eingeben und eine Liste der klinischen Studien finden, für die sie in Frage kommen.

Akalin und sein Team haben das Tool kürzlich verbessert, indem sie ihm beigebracht haben, all die verschiedenen Abkürzungen zu erkennen, die für ein und denselben Biomarker stehen – Rechtschreibfehler inbegriffen. Dadurch können Patient*innen, die nach klinischen Studien suchen, noch sicherer sein, dass sie eine vollständige Liste der Studien erhalten. Die verbesserte Suchfunktion von Onconaut haben die Forschenden kürzlich ebenfalls in „npj Digital Medicine“ beschrieben.

Text: Gunjan Sinha

Weiterführende Informationen

- Mit KI die passende Krebstherapie finden

Literatur

Farieda Gaber, Maqsood Shaik, Fabio Allega, et al. (2025) “Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis,” npj Digital Medicine DOI:10.1038/s41746-025-01684-1