

Studie: KI-Sprachmodelle übertreffen Fachmediziner

Fachkongressteilnehmende treten im Wissenstest zur akuten Nierenschädigung gegen KI an

Sind KI-Systeme bereits in der Lage, medizinisches Fachwissen besser abzurufen als Ärztinnen und Ärzte? Forschende der Universität Marburg und des Universitätsklinikums Gießen und Marburg (UKGM) untersuchten, wie gut 13 der bekanntesten öffentlich verfügbaren KI-Sprachmodelle klinisches Wissen zur akuten Nierenschädigung in einer standardisierten Testsituation abrufen und anwenden können. Dafür verglich das Team um Dr. Philipp Russ diese Modelle mit 123 freiwilligen Teilnehmenden, darunter Medizinstudierende sowie Ärztinnen und Ärzte aus dem Bereich der Inneren Medizin. Die ärztlichen Proband*innen waren Teilnehmende des 131. Jahreskongress der Deutschen Gesellschaft für Innere Medizin (DGIM). Die Tagung ist eine der größten internistischen Fachkongresse Europas mit rund 9.000 Teilnehmenden. Sie fand im Mai 2025 in Wiesbaden statt. Über die Ergebnisse berichten die Forschenden um Dr. Philipp Russ und Prof. Dr. Ivica Grgic im Fachblatt „Scientific Reports“ (<https://www.nature.com/articles/s41598-026-46846-7>)

Vorteil Maschine

Beide Gruppen bearbeiteten denselben deutschsprachigen Wissenstest zu Nierenschädigungen mit zwei realistischen Patientenfällen und 15 Multiple-Choice-Fragen. Das Ergebnis fiel deutlich aus: Die getesteten Sprachmodelle beantworteten im Mittel 90 Prozent der Fragen korrekt, die Fachkongressteilnehmenden im Vergleich dazu lediglich 49 Prozent. Mehrere Modelle beantworteten sämtliche Fragen korrekt und benötigten dafür nur einen Bruchteil der Zeit, die die Teilnehmenden brauchten.

Vorteil Mensch

Die Studie zeigt damit, dass große Sprachmodelle leitlinienkonformes medizinisches Fachwissen in standardisierten Fragesituationen inzwischen sehr zuverlässig wiedergeben können. Zugleich betonen die Autoren, dass ein gutes Abschneiden im Wissenstest nicht bedeutet, dass diese Systeme eigenständig klinische Entscheidungen treffen können oder gar sollten. „Menschliches Urteilsvermögen und klinische Erfahrung bleiben entscheidend. Die Letztverantwortung für die Versorgung von Patientinnen und Patienten liegt weiterhin klar bei den behandelnden Ärztinnen und Ärzten“, unterstreicht der Marburger Nephrologe und KI-Experte Prof. Dr. med. Ivica Grgic.

Chance für den klinischen Alltag

Studienleiter Dr. Philipp Russ kommentiert die Ergebnisse: „Große Sprachmodelle können medizinisches Faktenwissen sehr schnell bereitstellen. Das ist eine Chance für den klinischen Alltag. Gleichzeitig haben sie klare Grenzen: Sie können unter anderem fehlerhafte Inhalte erzeugen, den Menschen nicht in seiner ganzen Komplexität erfassen und keine Empathie empfinden. Ein Sprachmodell sieht nicht, hört nicht und spürt nicht, worum es einem Menschen wirklich geht. Genau deshalb kann es ärztliches Handeln und klinisches Urteil nicht ersetzen. Richtig eingesetzt könnte es uns aber mehr Zeit für das geben, was Patientinnen und Patienten besonders brauchen: Aufmerksamkeit, Zuwendung und menschliche Nähe.“

Perspektive

Vor diesem Hintergrund erscheint KI im klinischen Kontext aus heutiger Perspektive in erster Linie als unterstützendes Werkzeug. Gleichzeitig entzieht sich ihre weitere Entwicklung angesichts der hohen Innovationsdynamik einer verlässlichen Prognose. Für viele Einsatzbereiche ist die empirische Grundlage bislang noch begrenzt. Offen bleibt, ob und in welchem Umfang zukünftige Systeme autonomere Funktionen übernehmen werden und inwieweit eine solche Entwicklung gesellschaftlich gewünscht und akzeptiert ist. Die Integration in die klinische Praxis ist daher als schrittweiser Prozess zu verstehen, der einer kontinuierlichen fachlichen, regulatorischen und ethischen Reflexion bedarf.

Originalpublikation: Scientific Reports
(2026) <https://www.nature.com/articles/s41598-026-46846-7>