Ludwig-Maximilians-Universität München

22.05.2026 – 09:00 Uhr

Medizinische Diagnosen: Wie KI-Erklärungen Ärztinnen und Ärzten helfen

München (ots) -

- Eine neue Studie zeigt, dass KI-Modelle wie ChatGPT die diagnostische Genauigkeit in der Radiologie verbessern können
- Nicht jede Form der KI-Hilfe ist dabei gleich hilfreich. Entscheidend ist, ob die Ärztinnen und Ärzte die Empfehlung nachvollziehen und kritisch prüfen können.
- Knappe Antworten oder reine Listen können dagegen Fehlvertrauen fördern.

Große Sprachmodelle wie ChatGPT werden zunehmend als Unterstützung in der Medizin diskutiert. Sie können Informationen zusammenfassen, Diagnosen vorschlagen und ihre Einschätzungen in einfacher Sprache begründen. Gerade darin liegt ein zentrales Versprechen solcher Systeme: Sie liefern nicht nur eine Diagnose, sondern können auch erklären, warum eine Diagnose zutreffend ist. Doch ob solche Erklärungen Ärztinnen und Ärzten tatsächlich helfen - und welche Form besonders nützlich ist - ist bislang unklar.

Ein Forschungsteam der LMU München, des LMU Klinikums, des Karlsruher Instituts für Technologie und der Universität Bayreuth hat nun untersucht, wie unterschiedliche Formen von KI-Erklärungen die diagnostische Genauigkeit in der Radiologie beeinflussen. In einem randomisierten Experiment beurteilten 101 Radiologinnen und Radiologen reale klinische Fälle mit radiologischen Bildern, etwa aus der Computertomographie (CT) oder der Magnetresonanztomographie (MRT), und sollten jeweils eine Diagnose als Freitext formulieren. "In der Radiologie geht es oft darum, komplexe Bildbefunde mit klinischen Informationen zusammenzuführen", sagt Boj Friedrich Hoppe vom LMU Klinikum. "Sprachmodelle können hier prinzipiell unterstützen. Unsere Studie zeigt aber, dass nicht jede Form von KI-Hilfe gleich hilfreich ist. Entscheidend ist, ob die Ärztinnen und Ärzte die Empfehlung nachvollziehen und kritisch prüfen können."

Diagnose allein reicht nicht

Die Teilnehmenden wurden zufällig vier Gruppen zugeteilt: eine arbeitete ohne KI-Unterstützung, drei weitere erhielten unterschiedliche Hinweise eines multimodalen Sprachmodells. Die KI gab entweder nur eine Diagnose, eine Differentialdiagnose oder eine schrittweise "Chain-of-Thought"-Erklärung aus. Letztere erläuterte Bildmerkmale, klinische Hinweise und Ausschlusskriterien nachvollziehbar und half Ärztinnen und Ärzten besonders dabei, die Empfehlung mit ihrem Fachwissen abzugleichen.

"Für die klinische Praxis ist es nicht ausreichend, wenn ein KI-System nur eine plausibel klingende Antwort gibt", sagt Hoppe. "Ärztinnen und Ärzte müssen nachvollziehen können, welche Hinweise für eine Diagnose sprechen und wo mögliche Unsicherheiten liegen."

Schrittweise Erklärungen verbessern die Genauigkeit

Die Studie zeigt: Radiologinnen und Radiologen erzielten die höchste diagnostische Genauigkeit mit schrittweisen KI-Erklärungen - die Trefferquote lag 12,2 Prozentpunkte über der Kontrollgruppe ohne KI. Einfache Diagnoseausgaben und Differentialdiagnosen schnitten schlechter ab. Besonders bei fehlerhaften KI-Vorschlägen folgten Teilnehmende der Differentialdiagnose häufiger, was auf Automationsbias hindeutet. Schritt-für-Schritt-Erklärungen halfen dagegen, korrekte Hinweise gezielter zu übernehmen und Fehler eher zu erkennen.

Die Ergebnisse legen nahe, dass nicht allein die Qualität der Diagnose entscheidend ist, sondern auch die Form der Erklärung. Schrittweise Begründungen machen die Argumentation des Modells sichtbarer und erleichtern den Abgleich mit dem eigenen Fachwissen.

Differentialdiagnosen sind in der Medizin wichtig. In der Interaktion mit Sprachmodellen können sie jedoch mehrere Diagnosen nennen und so den Eindruck erwecken, der diagnostische Raum sei bereits vollständig abgedeckt. Das kann dazu führen, dass Ärzte bei seltenen oder komplexen Fällen weniger über die genannten Diagnosen hinausdenken.

Bedeutung über die Medizin hinaus

Die Studie fokussiert sich zwar auf die Radiologie, ihre Ergebnisse reichen laut Stefan Feuerriegel, Professor an der LMU Munich School of Management und korrespondierender Autor der Studie, aber weit darüber hinaus. Systeme wie ChatGPT würden zunehmend für Entscheidungen im Alltag und Beruf genutzt. "Unsere Ergebnisse zeigen: Wer nicht nur nach einer Antwort fragt, sondern auch nach einer nachvollziehbaren Begründung, kann solche Systeme deutlich besser nutzen." Entscheidend sei daher die Art der Interaktion. Nutzerinnen und Nutzer sollten KI-Antworten aktiv prüfen. "Eine gute KI-Antwort ist nicht nur korrekt, sondern überprüfbar", so Feuerriegel.

Vorsicht vor überzeugend klingenden Fehlern

Die Forschenden betonen, dass Sprachmodelle Fehler machen können - sowohl bei Diagnosen als auch bei deren Begründung. Gerade schrittweise Erklärungen können helfen, Empfehlungen kritisch zu prüfen. Die Studie zeigt: KI verbessert die diagnostische Leistung vor allem dann, wenn ihre Vorschläge nachvollziehbar präsentiert werden. Knappe Antworten oder reine Listen können dagegen Fehlvertrauen fördern.

Publikation

Philipp Spitzer, Daniel Hendriks, Jan Rudolph, Sarah Schlaeger, Jens Ricke, Niklas Kühl, Boj Friedrich Hoppe & Stefan Feuerriegel: *The effect of medical explanations from large language models on diagnostic accuracy in radiology*. In: *npj Digital Medicine*, Volume 9, Article 33, 2026. https://www.nature.com/articles/s41746-026-02619-0

Kontakt

Prof. Stefan Feuerriegel

Head of Institute of Artificial Intelligence (AI) in Management, LMU

E-Mail: feuerriegel@lmu.de

Tel.: +491627246860

Pressekontakt:

Claudia Russo
Ludwig-Maximilians-Universität München
Leopoldstr. 3
80802 München

Phone: +49 (0) 89 2180-2706
E-Mail: Claudia.Russo@lmu.de