

Studie liefert erste Hinweise auf Potenzial von KI bei psychopathologischer Befunderhebung

Eine Studie unter Federführung des Zentralinstituts für Seelische Gesundheit liefert erste Hinweise darauf, dass große Sprachmodelle komplexe psychopathologische Befunde aus Transkripten psychiatrischer Interviews mit einer Genauigkeit erfassen können, die mit den Ergebnissen überwiegend junger Klinikerinnen und Kliniker vergleichbar ist. Untersucht wurden zehn Sprachmodelle und 108 klinisch tätige Fachkräfte aus drei psychiatrischen Kliniken. Die Ergebnisse zeigen das Potenzial für KI-gestützte Entscheidungshilfen, sie sprechen aber nicht für einen Ersatz ärztlicher Beurteilung. Vor einem möglichen klinischen Einsatz sind weitere Studien mit realen Patientinnen und Patienten erforderlich.

Vom psychiatrischen Gespräch zum strukturierten Befund

Psychiatrische Diagnostik beginnt meist mit einem Gespräch. Fachkräfte erfassen dabei systematisch die psychische Verfassung eines Menschen. Aus häufig mehrdeutigen und unstrukturierten Informationen entsteht so ein psychopathologischer Befund, der die wichtigste Grundlage für Diagnostik und Therapie darstellt. Genau diesen Arbeitsschritt untersuchten die Forschenden. Während sich viele bisherige KI-Studien auf Diagnosen oder standardisierte Fragebögen konzentrieren, untersucht die vorliegende Arbeit erstmals systematisch, wie große Sprachmodelle bei der komplexen psychopathologischen Befunderhebung im Vergleich zu klinisch tätigen Fachkräften abschneiden. Die Ergebnisse wurden in der internationalen Fachzeitschrift npj Digital Medicine veröffentlicht.

Zehn Sprachmodelle im Vergleich mit 108 klinisch Tätigen

Für die Studie bewerteten zehn große Sprachmodelle die Transkripte dreier simulierter psychiatrischer Interviews zu Depression, Manie und Schizophrenie. Ihre Aufgabe bestand darin, sämtliche 100 Merkmale des AMDP-Systems einzuschätzen. Das AMDP-System ist ein in der deutschsprachigen Psychiatrie verbreitetes Verfahren, mit dem psychische Auffälligkeiten nach festen Kriterien systematisch beschrieben werden.

Zum Vergleich nahmen 108 klinisch tätige Ärztinnen, Ärzte sowie Psychologinnen und Psychologen aus drei psychiatrischen Kliniken dieselben Bewertungen vor. Die meisten Teilnehmenden befanden sich noch am Anfang ihrer klinischen Laufbahn. Während die Sprachmodelle ausschließlich die schriftlichen Transkripte erhielten, konnten die klinisch Tätigen die vollständigen Video- und Tonaufnahmen beurteilen, um eine möglichst realistische Bewertung zu ermöglichen.

Beste Sprachmodelle lagen im Bereich der klinischen Vergleichsgruppe

Die beiden leistungsstärksten Modelle, GPT-5.1 und Gemini-3-Pro-Preview, lagen insgesamt im Bereich der klinischen Vergleichsgruppe. Sie erreichten über alle drei Interviews hinweg eine durchschnittliche Genauigkeit von 72 Prozent, die klinische Vergleichsgruppe kam auf 68 Prozent. Die Werte beziehen sich ausschließlich auf die Einschätzung einzelner psychopathologischer Merkmale und nicht auf die Stellung einer psychiatrischen Diagnose. Aufgrund der geringen Anzahl

erfahrener Fachärztinnen und Fachärzte in der Stichprobe lässt die Studie keinen belastbaren Vergleich mit langjährig spezialisierten Psychiaterinnen und Psychiatern zu.

Klinische Fachkräfte und KI machen unterschiedliche Fehler

Besonders aufschlussreich war weniger die durchschnittliche Genauigkeit als die Art der Fehler. Die klinisch Tätigen neigten häufiger dazu, aus unvollständigen Informationen auf das Vorliegen oder Fehlen eines Merkmals zu schließen. Das Sprachmodell bewertete solche Merkmale dagegen häufiger als „nicht beurteilbar“. Besonders ausgeprägt war dieser Unterschied bei Symptomen, für deren Einschätzung nonverbale oder visuelle Hinweise wichtig sind. Die teilweise gegensätzlichen Fehlermuster legen nahe, dass sich menschliche klinische Erfahrung und eine streng am verfügbaren Text orientierte maschinelle Auswertung ergänzen könnten.

„Wir wollten bewusst nicht nur Diagnosen oder Fragebögen durch ein Sprachmodell auswerten lassen. Im psychiatrischen Alltag müssen aus komplexen und oft mehrdeutigen Gesprächen strukturierte psychopathologische Befunde erhoben werden“, sagt Erstautor Dr. Esra Lenz, Mitarbeiter am Hector Institut für Künstliche Intelligenz in der Psychiatrie (HITKIP) am Zentralinstitut für Seelische Gesundheit (ZI). „Unsere Ergebnisse liefern erste Hinweise darauf, dass KI diesen Prozess künftig unterstützen könnte. Sie ersetzt aber weder die klinische Erfahrung noch die unmittelbare Beurteilung durch eine Fachperson.“

Potenzial als zusätzliche Entscheidungshilfe

In einer nachträglichen Simulation untersuchten die Forschenden außerdem, ob ein Sprachmodell bei voneinander abweichenden klinischen Einschätzungen als zusätzliche Entscheidungshilfe dienen könnte. Wurde die Bewertung des Modells zur Auflösung solcher Uneinigkeiten herangezogen, waren die Ergebnisse genauer als bei einer zufälligen Auswahl zwischen den beiden klinischen Einschätzungen. Ähnliche Verbesserungen zeigten sich bei einer simulierten fachärztlichen Supervision. Diese Ergebnisse stammen jedoch ausschließlich aus einer statistischen Simulation und erlauben noch keine Aussage darüber, ob KI die Befunderhebung im klinischen Alltag tatsächlich verbessert.

Bestätigung mit realen Patientinnen und Patienten erforderlich

Die Forschenden betonen den explorativen Charakter der Untersuchung. Analysiert wurden lediglich drei simulierte Interviews, jeweils eines zu Depression, Manie und Schizophrenie. Reale Patientinnen und Patienten waren nicht beteiligt. Die Ergebnisse lassen sich daher weder auf psychiatrische Interviews allgemein noch auf einzelne Erkrankungen übertragen. Künftige Studien müssen prüfen, ob sich die Befunde bei einer größeren Anzahl unterschiedlicher Interviews, mit realen Patientinnen und Patienten und unter den Bedingungen des klinischen Alltags bestätigen. Dabei sind insbesondere hohe ethische und datenschutzrechtliche Anforderungen zu berücksichtigen.

„Die entscheidende Frage ist nicht, ob KI psychiatrische Fachkräfte ersetzt, sondern ob sie dort sinnvoll unterstützen kann, wo psychiatrische Diagnostik tatsächlich beginnt, nämlich im klinischen Gespräch und der psychopathologischen Befunderhebung“, sagt Prof. Dr. Emanuel Schwarz, Letztautor der Studie und Leiter des Hector Instituts für Künstliche Intelligenz in der Psychiatrie am ZI.

„Unsere Studie lenkt den Blick bewusst weg von Diagnosen und Fragebögen hin zum klinischen Gespräch und der psychopathologischen Befunderhebung. Bevor KI-Systeme in diesem Bereich eingesetzt werden können, müssen wir ihre Stärken und Grenzen unter realen Bedingungen

sorgfältig untersuchen“, ergänzt Dr. Tobias Gradinger, ebenfalls Letztautor und wissenschaftlicher Mitarbeiter am HITKIP.

###

Über das ZI

Das Zentralinstitut für Seelische Gesundheit (ZI) steht für international herausragende Forschung und wegweisende Behandlungskonzepte in Psychiatrie und Psychotherapie, Kinder- und Jugendpsychiatrie, Psychosomatik und Suchtmedizin. Die Kliniken des ZI gewährleisten die psychiatrische Versorgung der Mannheimer Bevölkerung. Psychisch kranke Menschen aller Altersstufen können am ZI auf fortschrittlichste, auf internationalem Wissensstand basierende Behandlungen vertrauen. Über psychische Erkrankungen aufzuklären, Verständnis für Betroffene zu schaffen und die Prävention zu stärken ist ein weiterer wichtiger Teil unserer Arbeit. In der psychiatrischen Forschung zählt das ZI zu den führenden Einrichtungen Europas und ist ein Standort des Deutschen Zentrums für Psychische Gesundheit (dzpg.org). Das ZI ist institutionell mit der Universität Heidelberg über gemeinsam berufene Professorinnen und Professoren der Medizinischen Fakultät Mannheim verbunden und Mitglied der Health + Life Science Alliance Heidelberg Mannheim (health-life-sciences.de).

Originalpublikation:

Lenz E, Naamanka J, Trabert W, Bottlender R, Malchow B, Meyer-Lindenberg A, Gradinger T, Schwarz E. Benchmarking large language models against practicing clinicians on psychopathological assessment. npj Digit Med. 2026;9(1):518. doi:10.1038/s41746-026-02852-7. <https://www.nature.com/articles/s41746-026-02852-7>