AUS DEM NETZWERK EVIDENZBASIERTE MEDIZIN

# Generative KI und evidenzbasierte Gesundheitsinformation

## Noch ein langer Weg

CHRISTOPH WILHELM UND DR. RER. NAT. FELIX G. REBITSCHEK IM AUFTRAG DES NETZWERKS EVIDENZBASIERTE MEDIZIN E. V. (WWW.EBM-NETZWERK.DE)

### AUFKLÄRUNG DURCH SPRACHMODELLE – EINE ILLUSION?

Die Nutzung generativer Sprachmodelle wie *ChatGPT* von *OpenAI, Gemini* von *Google* und *Le Chat* von *Mistral* AI nimmt auch im gesundheitsbezogenen Kontext rasant zu. Immer mehr Patientinnen und Patienten suchen mit ihrer Hilfe Informationen zu Prävention wie Behandlung. Dabei wurden Large Language Models (LLMs) ursprünglich gar nicht für die Suche nach Fakten konzipiert. Die Auflösung des Unterschieds zwischen Websuchen und LLMs lässt sich jedoch bereits bei Google beobachten. Die Stärken von LLMs liegen in der Textarbeit und der Persuasion [1]. Könnten sie dennoch den Anspruch erfüllen, evidenzbasierte, verständliche und ausgewogene Gesundheitsinformationen bereitzustellen? Eine Untersuchung von 2024er-Versionen von *ChatGPT, Gemini* und *Le Chat* liefert hierzu ernüchternde Ergebnisse.

### DER BEFUND – INFORMIERTES PROMPTEN WIRKT IN GRENZEN

Ein Forschungsteam um Dr. Felix Rebitschek vom Harding-Zentrum für Risikokompetenz in Potsdam hat im Frühjahr 2025 eine methodisch aufwendige Doppelstudie veröffentlicht. Beide untersuchten, inwieweit LLMs fähig sind, evidenzbasierte Informationen zu zwei Früherkennungsthemen bereitzustellen. Der thematische Fokus lag dabei auf der Krebsfrüherkennung, da es sich hierbei um weit verbreitete, medizinisch und gesellschaftlich besonders relevante Entscheidungsfelder handelt, die durch komplexe Nutzen-Risiko-Abwägungen sowie eine häufig verzerrte öffentliche Wahrnehmung gekennzeichnet sind [2, 3].

In der ersten, experimentell-analytischen Studie wurden die mithilfe eines strukturierten Prompting (Eingaben in ein Sprachmodell) generierten Inhalte von *ChatGPT, Gemini* und *Le Chat* systematisch hinsichtlich deren Übereinstimmung mit den Prinzipien evidenzbasierter Gesundheitskommunikation analysiert. Grundlage der Bewertung bildeten zwei Instrumente: die validierte Checkliste zur Bewertung der Qualität und Vertrauenswürdigkeit von Gesundheitsinformation - *MAPPinfo* [4] sowie das für diesen Zweck entwickelte Instrument *ebmNucleus*, das sich an der Leitlinie Evidenzbasierte Gesundheitsinfor-

mation [5] orientiert und ein stärkeres Gewicht auf entscheidungsrelevante Informationsanforderungen zur Nutzen-Risiko-Abwägung durch Laien legt.

Die Ergebnisse der Studie zeigten eindeutig, dass die Qualität der von den getesteten Sprachmodellen generierten Gesundheitsinformationen erheblich von der Informiertheit der formulierten Prompts abhing. Denn je gezielter die Prompts wichtige Prinzipien evidenzbasierter Kommunikation berücksichtigten, also zum Beispiel eine klare Abwägung von Nutzen und Schaden, die Nennung von Vergleichsgruppen oder die Angabe absoluter Zahlen, desto besser schnitten die Antworten der KI-Modelle in der Qualitätsbewertung ab. In der Analyse zeigte sich ein deutlicher Zusammenhang zwischen der Qualität der Prompts und der Qualität der Antworten: Bei *MAPPinfo* ergab sich ein mittlerer und bei *ebmNucleus* ein sehr großer Effekt. Das bedeutet, dass gut aufgebaute Prompts die Qualität der Antworten spürbar steigern können [6].

Gleichwohl blieben die absoluten Ergebnisse weit hinter dem fachlich erforderlichen Niveau zurück: Kein einziges der getesteten Sprachmodelle erreichte, selbst unter bestmöglichen Bedingungen, mehr als 50 Prozent der maximal möglichen Qualitätspunkte von *MAPPinfo*. Besonders auffällig waren dabei wiederkehrende Defizite in zentralen Bereichen: Konkrete Zahlen zu Nutzen und Schaden der jeweiligen Vorsorgeuntersuchungen wurden kaum genannt, Risikodifferenzierungen blieben unvollständig, und Hinweise auf Quellen, die Evidenzlage oder bestehende Unsicherheiten fehlten nahezu durchgängig [6].

Die zweite, randomisiert-kontrollierte Studie ergänzte die zuvor beschriebene systematische Analyse um eine nutzerzentrierte Perspektive. Hier wurden 300 englischsprachige medizinische Laien über Prolific.co rekrutiert und zufällig einem der drei LLMs zugewiesen. Die Teilnehmenden erhielten die Aufgabe, eigene Fragen zur Mammografie zur Brustkrebsfrüherkennung oder zum PSA-Test zur Prostatakrebsfrüherkennung zu formulieren und diese zu prompten. Die Hälfte der Gruppe erhielt einen Boost. Boosts sind niedrigschwellige verhaltenspsychologische Interventionen, die auf eine Kompetenzsteigerung für selbstbestimmtes Entscheiden abzielen [7]. Hierzu wurde konkret eine sogenannte OARS-Regel instruiert ohne weitere Übung oder Anwendungshilfe. Die OARS-Regel (Options, Advantages, Risks, Steadiness) fasst zentrale Aspekte informierter medizinischer Entscheidungen in kompakter und einprägsamer Weise zusammen und orientiert sich konzeptionell an der evidenzbasierten Kommunikationsstrategie *AskShareKnow* [8], wurde jedoch speziell für den LLM-Kontext durch die Autoren angepasst. Sie forderten die Teilnehmenden dazu auf, gezielt an die Optionen, möglichen Nutzen, Schaden und ihr Eintreten zu denken, also an Kernelemente der Entscheidung auf Basis evidenzbasierter Gesundheitskommunikation. Ziel der Studie war es, die Wirkung einer minimalen Intervention bezüglich des Promptings auf die Qualität der LLM-Antworten infolge dieses Promptings zu untersuchen.

Die Ergebnisse zeigten, dass die Bereitstellung der OARS-Regel die Qualität der von den Sprachmodellen generierten Informationen verbessern konnte. Dieser Effekt war unabhängig vom verwendeten Modell und zeigte sich sowohl im *MAPPinfo*-Score mit einem kleinen bis mittleren Effekt auch im *ebmNucleus*-Score mit einem kleinen Effekt. Insgesamt blieb die Qualität der Antworten jedoch weiterhin unter dem fachlich wünschenswerten Niveau [6].

Obwohl sich durch die verhaltensorientierte Intervention die Ergebnisse verbesserten, blieben sie in ihrer Gesamtheit jedoch ernüchternd. Die Korrektheit der Antworten lag weiterhin deutlich unter dem fach-

lich wünschenswerten Niveau. Besonders häufig fehlten weiterhin Angaben zur Bezugsgröße von Zahlen (etwa absolute statt relative Risikoangaben), Hinweise auf die Zielgruppe der Aussagen oder auf die Herkunft und Qualität der zugrunde gelegten Evidenz. Auch die Darstellung möglicher Unsicherheiten, Nebenwirkungen oder Schäden im Zusammenhang mit den jeweiligen Früherkennungsmaßnahmen blieb in der Mehrzahl der Antworten lückenhaft [6].

Die Studie verdeutlicht damit zweierlei: Einerseits lässt sich durch niedrigschwellige, verhaltensorientierte Interventionen wie die OARS-Regel die Qualität der generierten Inhalte verbessern. Andererseits zeigen die weiterhin bestehenden Defizite, dass solche Maßnahmen allein nicht ausreichen, um fundamentale Grenzen generativer Sprachmodelle in der evidenzbasierten Gesundheitskommunikation zu überwinden.

Sprachmodelle differenzieren bislang kaum zwischen relevanten und irrelevanten Informationen. Sie liefern selten überprüfbare bzw. die relevanten Quellen oder Hinweise auf Leitlinien – und zentrale Elemente evidenzbasierter Patienteninformationen, wie z. B. die Angabe von absoluten Risikodifferenzen, Angaben zu Überdiagnosen oder potenziellen Folgeschäden, fehlen weitgehend. Auch Kontextinformationen, etwa zu Zielgruppen, Altersbereichen oder Früherkennungsintervallen, werden meist nicht berücksichtigt, was insbesondere bei vulnerablen Nutzergruppen zu Verunsicherung führen kann. Selbst wenn die Nutzerin oder der Nutzer explizit nach dem spezifischen Nutzen und Schaden fragte, lieferten die Modelle überwiegend vage, häufig verzerrte oder selektiv positive Darstellungen [6].

## KONSEQUENZEN FÜR DIE ÄRZTLICHE PRAXIS

Für die ärztliche Praxis bedeutet dies, dass auch vermeintlich „besser formulierte“ Nutzeranfragen keine Garantie für sachlich korrekte, vollständige oder ausgewogene Inhalte bieten. Da generative Sprachmodelle dennoch zunehmend von Patientinnen und Patienten genutzt werden, sehen sich Ärztinnen und Ärzte immer häufiger in der Rolle, fehlerhafte oder lückenhafte KI-Informationen einzuordnen und zu korrigieren.

Hieraus erwächst ein doppelter Handlungsauftrag: Einerseits braucht es klare Kommunikation in der Sprechstunde über die Möglichkeiten und Grenzen dieser Informationstechnologien. Andererseits sollte das medizinische Fachpersonal selbst eine gewisse Kompetenz im Umgang mit generativen Sprachmodellen entwickeln, um ihre Stärken und Schwächen differenziert bewerten zu können. Denkbar wäre auch die Entwicklung medizinisch validierter Interfaces, die auf geprüfte Evidenzdatenbanken zugreifen und strukturiert aufgebaut sind, wie etwa über eine Integration von Leitlinieninhalten oder Cochrane-Datenbanken.

Langfristig könnten generative Sprachmodelle durchaus zur Verbesserung bzw. Steigerung der Gesundheitskompetenz beitragen. Etwa durch personalisierte Aufklärung oder niedrigschwellige Erläuterung medizinischer Konzepte. Voraussetzung dafür wäre jedoch eine technische und inhaltliche Qualitätssicherung, die derzeit nicht gegeben ist. Auch auf Patientenseite bleibt deshalb heute die Förderung gesundheitsbezogener Informationskompetenz ebenso notwendig wie eine kritische Reflexion der systematischen Grenzen dieser Technologien.

## FAZIT

Generative KI-Systeme wie *ChatGPT, Gemini* oder *Le Chat* wollen und können bislang nicht den Anspruch erfüllen, evidenzbasierte Gesundheitsinformationen zu liefern. Ihre Inhalte bleiben trotz gezielter Steuerung durch Prompts deutlich hinter den Anforderungen zurück, die sich aus der Leitlinie evidenzbasierte

Gesundheitsinformation bzw. *MAPPinfo* an Qualität, Transparenz und Risikokommunikation ergeben. Auch wenn sich ihre Nutzung durch Patientinnen und Patienten nicht verhindern lässt, sollten Ärztinnen und Ärzte sich der begrenzten Aussagekraft bewusst sein und aktiv aufklären. Ein künftiger, sicherer Einsatz in der Medizin ist denkbar, aber nur unter der Voraussetzung klarer Qualitätsstandards, unabhängiger Überprüfung und Integration in evidenzbasierte Entscheidungsprozesse. ■

**CHRISTOPH WILHELM**
**M.Sc. Gesundheitswissenschaften**
**Wissenschaftlicher Mitarbeiter**
**Harding-Zentrum für Risikokompetenz /**
**Fakultät für Gesundheitswissenschaften**
**Universität Potsdam**
**christoph.wilhelm@uni-potsdam.de**

**DR. FELIX G. REBITSCHEK**
**Wissenschaftlicher Leiter und Geschäftsführer**
**Harding-Zentrum für Risikokompetenz /**
**Fakultät für Gesundheitswissenschaften**
**Universität Potsdam**

Literaturverzeichnis
1. Costello TH, Pennycook G, Rand DG. Durably reducing conspiracy beliefs through dialogues with AI. Science. 2024;385(6714):eadq1814. Epub 20240913. doi: 10.1126/science.adq1814. PubMed PMID: 39264999.
2. Wegwarth O, Widschwendter M, Cibula D, Sundström K, Portuesi R, Lein I, et al. What do European women know about their female cancer risks and cancer screening? A cross-sectional online intervention survey in 5 European countries. BMJ Open. 2018;8. doi: 10.1136/bmjopen-2018-023789.
3. Gigerenzer G, Mata J, Frank R. Public knowledge of benefits of breast and prostate cancer screening in Europe. J Natl Cancer Inst. 2009;101(17):1216-20. Epub 20090811. doi: 10.1093/jnci/djp237. PubMed PMID: 19671770; PubMed Central PMCID: PMC2736294.
4. Kasper J, Luhnen J, Hinneburg J, Siebenhofer A, Posch N, Berger-Hoger B, et al. MAPPinfo - mapping quality of health information: Validation study of an assessment instrument. PLoS One. 2023;18(10):e0290027. Epub 20231023. doi: 10.1371/journal.pone.0290027. PubMed PMID: 37871040; PubMed Central PMCID: PMC10593225.
5. Lühnen J, Albrecht M, Mühlhauser I, Steckelberg A. Leitlinie evidenzbasierte Gesundheitsinformation. Hamburg;2017 [28.09.2025]. Abrufbar unter: http://www.leitlinie-gesundheitsinformation.de/.
6. Rebitschek FG, Carella A, Kohlrausch-Pazin S, Zitzmann M, Steckelberg A, Wilhelm C. Evaluating evidence-based health information from generative AI using a cross-sectional study with laypeople seeking screening information. NPJ Digit Med. 2025;8(1):343. Epub 20250609. doi: 10.1038/s41746-025-01752-6. PubMed PMID: 40490558; PubMed Central PMCID: PMC12149300.
7. Herzog SM, Hertwig R. Boosting: Empowering Citizens with Behavioral Science. Annu Rev Psychol. 2025;76(1):851-81. Epub 20241203. doi: 10.1146/annurev-psych-020924-124753. PubMed PMID: 39413154.
8. Shepherd HL, Barratt A, Jones A, Bateson D, Carey K, Trevena LJ, et al. Can consumers learn to ask three questions to improve shared decision making? A feasibility study of the ASK (AskShareKnow) Patient-Clinician Communication Model((R)) intervention in a primary health-care setting. Health Expect. 2016;19(5):1160-8. Epub 20150914. doi: 10.1111/hex.12409. PubMed PMID: 26364752; PubMed Central PMCID: PMC5152736.

EbM Netzwerk