Nutzung generativer KI-Modelle in der Inneren Medizin

Ivica Grgic[1,2] · Martin Möckel[3] · Sebastian Kuhn[4] · Mirja Mittermaier[5] · Johannes Knitza[4] · Georg Ertl[6,7] · Claus Vogelmeier[8] · Martin C. Hirsch[2]

[1] Zentrum für Innere Medizin – SP Nephrologie, Universitätsklinikum Gießen und Marburg (UKGM), Philipps-Universität Marburg, Marburg, Deutschland
[2] Institut für Künstliche Intelligenz in der Medizin, Philipps-Universität Marburg, Marburg, Deutschland
[3] Notfall- und Akutmedizin, Zentrale Notaufnahmen mit Chest Pain Units, Charité - Universitätsmedizin Berlin, Berlin, Deutschland
[4] Institut für Digitale Medizin, Universitätsklinikum Gießen und Marburg, Philipps-Universität Marburg, Marburg, Deutschland
[5] Fächerverbund für Infektiologie, Pneumologie und Intensivmedizin, Charité - Universitätsmedizin Berlin, Berlin, Deutschland
[6] Universitätsklinikum Würzburg, Deutsches Zentrum für Herzinsuffizienz, Würzburg, Deutschland
[7] Deutsche Gesellschaft für Innere Medizin (DGIM), Wiesbaden, Deutschland
[8] Klinik für Innere Medizin mit Schwerpunkt Pneumologie, Intensiv- und Schlafmedizin, Universitätsklinikum Gießen und Marburg (UKGM), Standort Marburg, Marburg, Deutschland

DGIM Deutsche Gesellschaft für Innere Medizin

Redaktion
Prof. Dr. Georg Ertl (v. i. S. d. P.)

Korrespondenzadresse
Deutsche Gesellschaft für Innere Medizin e. V. (DGIM)
Irenenstraße 1
65189 Wiesbaden
Tel. 0611/205 8040-0 Fax 0611/205 8040-46
info@dgim.de

Vorwort

„Große Sprachmodelle" (Large Language Models, LLMs) und verwandte Systeme generativer künstlicher Intelligenz (KI) wie Large Multimodal Models (LMMs) haben in der Medizin schon heute erhebliche Bedeutung erlangt. Sie haben das Potenzial, Informationen schneller zugänglich zu machen, Texte zu strukturieren und administrative Prozesse zu unterstützen und damit ärztliche Alltagsarbeit spürbar zu entlasten. Darüber hinaus zeigen sie erhebliches Potenzial bei der Diagnosefindung und Therapieplanung.

Sogenannte *Ambient Scribes*, die klinische Gespräche erfassen und daraus zeitnah strukturierte Dokumentationsentwürfe generieren, wurden entscheidend verbessert. Allerdings variieren Nutzen und Leistungsfähigkeit erheblich, und inhaltliche Ungenauigkeiten treten weiterhin auf, insbesondere in komplexen Situationen, wie etwa in der Notaufnahme. Generative KI-Systeme werden in der Praxis häufig noch unkritisch eingesetzt, und es fehlen einheitliche Standards. Daraus ergeben sich zentrale Fragen zur Verlässlichkeit und Nachvollziehbarkeit der Ausgaben, zu Datenschutz und Informationssicherheit, zu ethischen Implikationen sowie zur rechtlichen Verantwortung. Risiken entstehen insbesondere dann, wenn die Funktionsweise und Grenzen der Systeme nicht ausreichend verstanden werden.

Zugleich zeichnet sich bereits die nächste Entwicklungsstufe ab: Systeme mit agentischen Eigenschaften, die Koordination, Aufgabendelegation, das Zusammenwirken mehrerer (digitaler) Agenten und längere autonome Laufzeiten umfassen. Für die Medizin bedeutet dies, dass mit zunehmender Agentizität auch die Risiken steigen. Es braucht daher eine fortlaufende Weiterentwicklung der Orientierung und Rahmenbedingungen für KI.

Die Deutsche Gesellschaft für Innere Medizin (DGIM) hat in ihrem Strategiepapier „Die Innere Medizin 2025/2030" (Ertl et al. 2025) die digitale Transformation der Medizin und die Nutzung von KI für die medizinische Forschung und Patientenversorgung als einen ihrer zentralen Schwerpunkte definiert. „Wir sind mitten in einem Wandel, der vergleichbar ist mit dem Einzug der Naturwissenschaften in die Medizin im 19. Jahrhundert." Es weist auch darauf hin, dass „jenseits der jetzt gängigen KI (…) im medizinischen und wissenschaftlichen Einsatz die Nutzung der Gesamtheit der medizinisch wissenschaftlichen Daten erst das gesamte Potenzial von KI ausschöpfen" wird. Der vorliegende Beitrag knüpft hieran an und unterstreicht den Anspruch der Fachgesellschaft, die Entwicklung, Evaluierung und verantwortungsvolle Integration dieser Technologien aktiv mitzugestalten. Er greift zentrale Leitgedanken des Strategiepapiers auf, die „Risiken und ungeklärte Fragen" definieren und insbesondere die konsequente Nutzung der Potenziale von KI bei gleichzeitiger Wahrung ärztlicher Verantwortung, Entscheidungsqualität und der Arzt-Patienten-Beziehung betonen. Nicht zuletzt wird „die DGIM (…) Kriterien dafür entwickeln, was in Zukunft in der Inneren Medizin der KI überlassen werden kann".

Vor diesem Hintergrund sieht die DGIM aktuell die Notwendigkeit, Orientierung zu geben, und versteht dies auch künftig als eine ihrer zentralen Aufgaben als wissenschaftliche Fachgesellschaft. Der vorliegende Beitrag soll hierbei einen Anfang machen und einen reflektierten, sicheren und verantwortungsvollen Umgang mit großen Sprachmodellen und verwandten generativen KI-Systemen im klinischen Alltag unterstützen. Er benennt Potenziale, grenzt sinnvolle Einsatzfelder ab und formuliert Bedingungen, unter denen Anwendungen patientenzentriert, rechts- und datenschutzkonform sowie qualitätsgesichert genutzt werden können.

1. Einleitung und Zielsetzung

Large Language Modelle (LLMs) sind großskalige neuronale Sprachmodelle, die auf umfangreichen Datenmengen trainiert werden und Sprache auf Grundlage erlernter statistischer Muster verarbeiten und generieren können. Es handelt sich nicht um „Verstehen"

Innere Medizin 2026 · 67:1235–1246
https://doi.org/10.1007/s00108-026-02183-5
© The Author(s), under exclusive licence to Springer Medizin Verlag GmbH, ein Teil von Springer Nature 2026

der Texte, also um aktive Sinnentnahme zum Aufbau einer modellhaften Repräsentation der Welt.

Im klinischen Umfeld werden zunehmend auch weiterentwickelte generative Systeme relevant, die neben Text weitere Modalitäten wie Bilder, Audiodaten oder Videos verarbeiten und erzeugen können, sogenannte Large Multimodal Models (LMMs). Beide Systemklassen haben sich innerhalb kurzer Zeit zu Werkzeugen entwickelt, die im klinischen Alltag real genutzt werden – oft zunächst informell, etwa zur Textproduktion, Informationsaufbereitung oder Unterstützung bei organisatorischen Aufgaben (Thirunavukarasu et al. 2023).

Für die Innere Medizin ist diese Entwicklung besonders relevant, da ihr Fachgebiet häufig komplexe Krankheitsbilder umfasst, die einerseits große Informationsmengen in diagnostischen Daten und andererseits in multiplen Leitlinien dokumentierte Evidenz verbinden. Entsprechend berührt ihr Einsatz sowohl die Versorgungspraxis als auch die medizinische Forschung und zukünftige Aus- und Weiterbildung (Singhal et al. 2023). Internistinnen und Internisten werden sich auch darauf einstellen müssen, dass Patientinnen und Patienten generative KI-basierte Systeme nutzen.

Mit der weiteren Verbreitung von generativen KI-Systemen wächst die Hoffnung auf Entlastung in der Patientenversorgung, eine bessere Zugänglichkeit zu Informationen und Unterstützung bei komplexen Entscheidungen (Solmonovich et al. 2025). Dem stehen offene Fragen gegenüber: Welche Anwendungen sind fachlich sinnvoll und verantwortbar? Wo liegen die Grenzen der Verlässlichkeit? Wie lassen sich Datenschutz, ärztliche Sorgfaltspflichten und klinische Verantwortung mit neuen KI-gestützten Arbeitsabläufen vereinbaren? Und wie verhalten sich die medizinischen Fähigkeiten einer KI zu den ethischen Selbsterklärungen ihrer Hersteller? LLMs können Fehler mit hoher Eloquenz erzeugen und ihre Leistungsfähigkeit hängt stark vom Anwendungskontext, von den Eingaben und von den organisatorischen Rahmenbedingungen ab (Bedi et al. 2025). Hinzu kommt, dass die wissenschaftliche Evidenz zu Nutzen und Fehleranfälligkeit von LLMs trotz der rasanten Publikationsdynamik bislang begrenzt ist. Eine aktuelle systematische Auswertung von über 4600 klinischen Studien zu LLMs zeigte, dass nur ein kleiner Teil der Arbeiten auf realen Patientendaten basiert und prospektive randomisierte Studien bislang selten sind (Chen et al. 2026). Negative Studien machen auf die Komplexität erforderlicher Studiendesigns aufmerksam (Hautz et al. 2025). Ein systematisches Review kommt zu der Schlussfolgerung, dass insbesondere die ärztliche Verantwortung bei der medizinischen Entscheidung außer Frage steht, die Entscheidungsfindung aber eine wertvolle Unterstützung durch LLMs erhalten kann (Mohajer-Bastami et al. 2025).

Mit der technologischen Weiterentwicklung wird zudem eine begriffliche Präzisierung erforderlich. Während LLMs und LMMs primär auf die Verarbeitung und Generierung von Inhalten ausgerichtet sind, werden im klinischen Umfeld zunehmend auch Systeme mit agentischen Eigenschaften diskutiert. Dabei ist zwischen einzelnen KI-Agenten und agentischer KI zu unterscheiden: KI-Agenten bezeichnen Einzelsysteme, die mit einem definierten Maß an Autonomie konkrete Ziele verfolgen, während agentische KI komplexere Architekturen beschreibt, in denen Aufgaben über mehrere Schritte, unter Nutzung externer Werkzeuge und in koordinierten Teilsystemen geplant und bearbeitet werden (Organization for Economic Co-operation and Development (OECD) Artificial Intelligence Papers, No. 56, 2026). Generative KI bildet für beide Formen die technische Grundlage. Für die klinische Praxis ist diese Unterscheidung von Bedeutung, da mit wachsender digitaler Agentizität nicht nur das Potenzial, sondern auch die Anforderungen an Validierung, menschliche Aufsicht, Protokollierung und Governance steigen.

Der Einsatz generativer KI in der Medizin ist nicht nur fachlich und ethisch, sondern zunehmend auch regulatorisch einzuordnen. Mit dem EU AI Act wird der europäische Rechtsrahmen für Entwicklung, Bereitstellung und Nutzung von KI-Systemen schrittweise konkretisiert. Zentrale Bestimmungen, darunter zu Verboten, KI-Kompetenz und General-Purpose-AI-Modellen, gelten bereits. Für den klinischen Alltag besonders relevant sind Anforderungen an KI-Kompetenz, Transparenz und menschliche Aufsicht sowie, je nach Zweckbestimmung und Risikoprofil, eine weitergehende regulatorische Einordnung einzelner Anwendungen.

Ziel des Papiers ist es daher, einen Rahmen zu definieren, in dem generative KI-Systeme genau dort zum Einsatz kommen, wo sie die Qualität und Effizienz der Versorgung tatsächlich verbessern, ohne dass Patientensicherheit, Datenschutz oder professionelle Standards gefährdet werden. Im Mittelpunkt stehen eine informierte Nutzung mit klaren Anwendungsgrenzen, überprüfbare Qualitätsanforderungen und die konsequente Einordnung von KI-Ausgaben als unterstützende Vorschläge, die ärztlich validiert werden müssen.

Abschließend skizziert das Papier, wie der Umgang mit KI-Anwendungen künftig bereits im Studium vermittelt werden sollte, damit kommende Ärztinnen- und Ärztegenerationen im neuen technologischen Umfeld souverän handlungsfähig sein können.

2. Potenziale von generativer KI im Klinikalltag

2.1 Dokumentations- und Verwaltungsentlastung

Die Entlastung von Dokumentations- und Verwaltungsaufgaben stellt derzeit eines der offensichtlichsten und zugleich am wenigsten kontroversen Einsatzfelder von LLMs im klinischen Alltag dar. LLM-basierte Schreibassistenten sind in der Lage, ärztliche Texte zu strukturieren, zusammenzufassen und formal zu vereinheitlichen – etwa bei der Befundung, der Verlaufsdokumentation oder bei Arztbriefen (Williams et al. 2025). Damit adressieren sie einen Bereich, der einen erheblichen Anteil der ärztlichen Arbeitszeit bindet, ohne unmittelbar zur patientennahen Versorgung beizutragen (s. **Abb. 1**).

Eine besondere Ausprägung dieser Entwicklung sind die sogenannten *Ambient Scribes*, die Inhalte aus dem Arzt-Patienten-Gespräch erfassen und automatisiert in strukturierte Dokumentationsentwürfe überführen (Lawrence et al. 2025). Richtig eingesetzt, können sie nicht nur den Dokumentationsaufwand reduzieren, sondern sogar dazu beitragen, die ärztliche Aufmerksamkeit stärker auf das Gespräch selbst zu richten. Die besondere Herausforderung liegt dabei in der korrekten Erfassung und Kontextzuordnung gesprochener Inhalte in Echtzeit oder unmittelbar im Anschluss an das Gespräch. Davon zu unterscheiden ist die KI-gestützte Arztbriefschreibung, die typischerweise auf Informationen aus einem

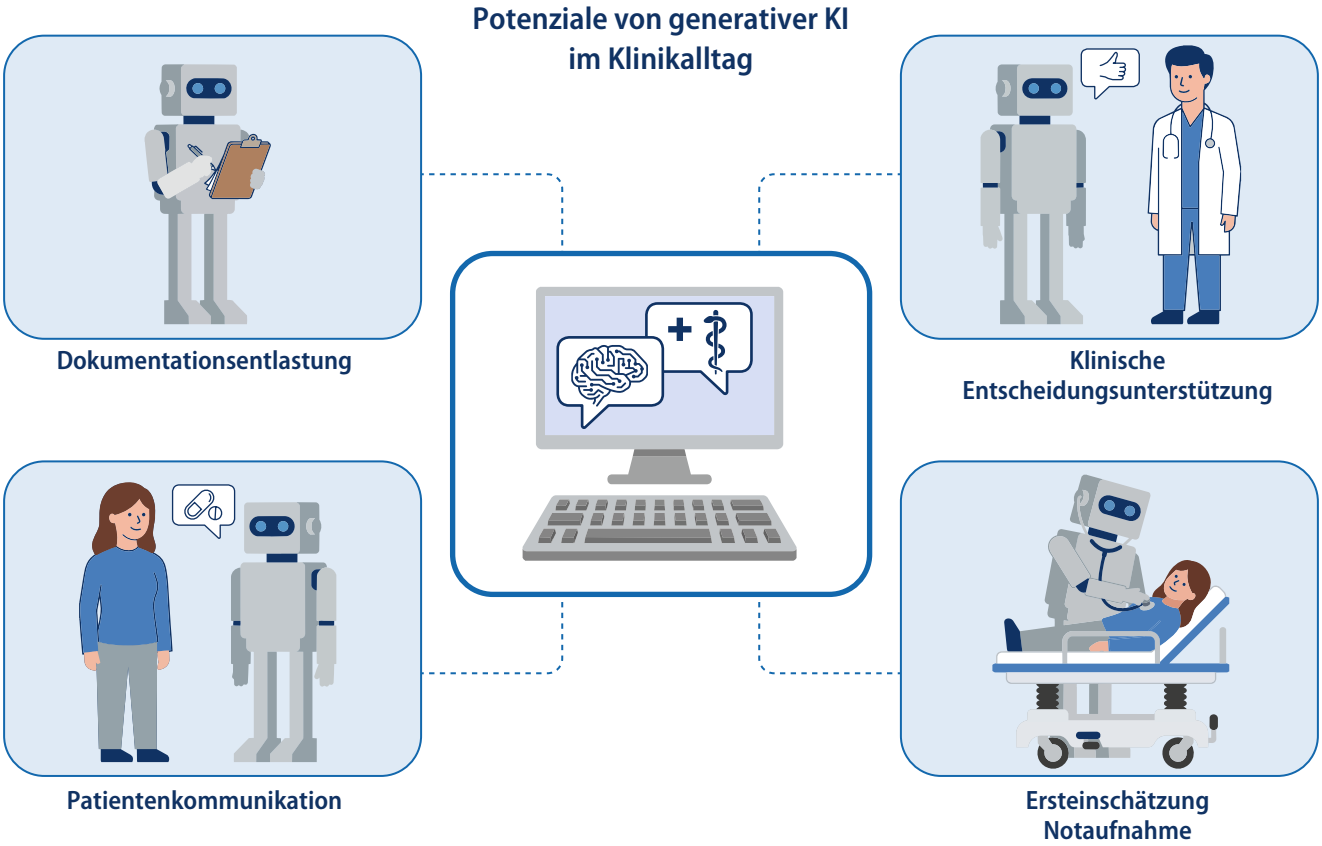


Abb. 1 ▲ Potenziale generativer KI im klinischen Alltag. Schematische Darstellung zentraler Einsatzfelder generativer KI in der internistischen Versorgung. Im Zentrum steht das LLM als unterstützendes, wissensbasiertes Assistenzsystem, das in verschiedene klinische Arbeitsbereiche eingebunden ist. *Oben links*: Entlastung bei der Dokumentation – automatisierte Strukturierung, Zusammenfassung und formale Vereinheitlichung medizinischer Texte sowie Unterstützung durch Ambient Scribe-Systeme, die Arzt-Patienten-Gespräche automatisch erfassen und dokumentieren. *Oben rechts*: Klinische Entscheidungsunterstützung – Vorschlagen und Strukturieren von Differenzialdiagnosen, Aufbereitung von Therapieoptionen und leitlinienorientierte Wissenssynthese im Sinne einer kognitiven Assistenz. *Unten links*: Patientenkommunikation – Übersetzung komplexer medizinischer Inhalte in verständliche, patientenzentrierte Informationen zur Förderung der Gesundheitskompetenz und der partizipativen Entscheidungsfindung (Shared Decision-Making). *Unten rechts*: KI-gestützte Ersteinschätzung in der Notaufnahme – Unterstützung bei der Priorisierung nach Behandlungsdringlichkeit und bei der Erkennung potenziell kritischer Konstellationen auf Basis verlässlich erhobener Eingangsdaten (klinische Befunde, Symptome, Vitalparameter, Anamnese)

längeren stationären Behandlungsverlauf mit mehreren Dokumentationsquellen und Beteiligten beruht. Hier stehen inhaltliche Integration, Vollständigkeit und die abschließende, klar zugeordnete ärztliche Verantwortung für den Behandlungsprozess im Vordergrund. Für diese Anwendungen liegen inzwischen erste belastbare Daten dafür vor, dass sich der Zeitaufwand für Dokumentation reduzieren lässt (Song et al. 2025). Erste experimentelle Studien zu konversationsbasierten medizinischen KI-Systemen zeigen zudem, dass solche Modelle in strukturierten Dialogsituationen diagnostisch relevante Informationen erfassen und aufbereiten können (Tu et al. 2025).

Neben der unmittelbaren Zeitersparnis können standardisierte und konsistente Dokumentationsprozesse auch die interne Kommunikation verbessern, etwa zwischen Stationen, Fachabteilungen oder im Übergang zwischen ambulanter und stationärer Versorgung. Auch bei der Erstellung strukturierter Entwürfe für Korrespondenz und Anträge (z. B. auf Off-Label-Therapien), bei sozialmedizinischen Stellungnahmen oder der allgemeinen fachlichen Kommunikation können LLMs unterstützend wirken (Tripathi et al. 2024). Allerdings besteht auch das Risiko, dass solche Systeme nicht zu mehr Freiräumen für genuin ärztliche Tätigkeiten führen, sondern für eine weitere Verdichtung von Abläufen genutzt werden. Hier droht dann eine unreflektierte Anwendung bei weiterhin voller ärztlicher Verantwortung mit persönlichen Risiken für Ärztinnen und Ärzte und medizinischen Risiken für die Patientinnen und Patienten.

2.2 Unterstützung bei klinischen Entscheidungen

Generative KI-Systeme eröffnen die Möglichkeit, ärztliche Entscheidungsprozesse als kognitive Assistenzsysteme zu unterstützen (McDuff et al. 2025). In dieser Rolle fungieren sie nicht als autonome Entscheidungsträger, sondern als Instrumente der digitalen Entscheidungsvorbereitung: Sie können helfen, komplexe Informationen zu ordnen, Differenzialdiagnosen vorzuschlagen und zu strukturieren oder Therapieoptionen systematisch gegenüberzustellen (Labinsky et al. 2024). Darüber hinaus können sie bei der Reflexion medizinethisch komplexer Entscheidungssituationen unterstützen, indem sie ethische Prinzipien, klinische Handlungsoptionen und mögliche Konsequenzen gegenüberstellen (Singhal et al. 2025).

Ein wesentlicher Vorteil generativer KI liegt im schnellen Zugriff auf große Textmengen aus Leitlinien, Reviews und anderen Wissensquellen, die synthetisiert und in den gewünschten klinischen Kontext eingeordnet werden (Russ et al. 2026). Dies ist im klinischen Alltag hochrelevant, in dem klassische Recherchestrategien häufig zu zeitaufwendig sind. Ähnliches gilt für die strukturierte Aufbereitung komplexer Fälle: Generative KI kann helfen, relevante Informationen zu ordnen, Zusammenhänge sichtbar zu machen und Entscheidungsprozesse vorzubereiten, ohne dabei selbst eine klinische Bewertung vorzunehmen. Beide Funktionen sind in ihrer Wirkung auf die Informationsaufbereitung begrenzt und tangieren die eigentliche ärztliche Entscheidung nicht unmittelbar, was ihre regulatorische Einordnung vergleichsweise unkompliziert macht.

Demgegenüber sind diagnostische Unterstützung, Therapieabwägung und ethische Reflexion mit einem deutlich höheren Risiko verbunden. Konkrete Anwendungsfelder umfassen unter anderem die Prüfung von Arzneimittelwechselwirkungen, die Unterstützung bei diagnostischen Überlegungen sowie die strukturierte Aufbereitung von Therapieoptionen unter Berücksichtigung von Leitlinien und Komorbiditäten. Gerade bei den für die Innere Medizin typischen Multimorbiditäten und bei komplexen oder seltenen Krankheitsbildern kann generative KI dazu beitragen, relevante Aspekte sichtbar zu machen (Chase et al. 2026). Bei konkreten klinischen Empfehlungen sind die Anforderungen an die Validität, Nachvollziehbarkeit und regulatorische Zulässigkeit der KI besonders hoch. Diagnostische Vorschläge und Therapieempfehlungen dürfen nicht unkritisch übernommen werden. LLM-Ausgaben sind zwingend als vorläufig zu verstehen und bedürfen einer expliziten ärztlichen Prüfung.

Für die ethische Reflexion gilt dies in besonderem Maße: Generative KI-Systeme können helfen, relevante Prinzipien und Handlungsoptionen zu strukturieren, ersetzen jedoch keinesfalls das ärztliche Urteilsvermögen in individuellen, oft biografisch und kulturell geprägten Entscheidungssituationen. Der Einsatz von LLMs zur Unterstützung ethischer Überlegungen setzt ein hohes Maß an Reflexionskompetenz auf Seiten der Anwendenden voraus und darf nicht dazu führen, dass Verantwortung an ein technisches System delegiert wird.

Eine aktuelle, in *Science* publizierte Untersuchung verglich ein *Reasoning*-Modell, also ein generatives KI-System, das Aufgaben vor der Antwort schrittweise durcharbeitet, direkt mit mehreren Hundert Ärztinnen und Ärzten unterschiedlicher Erfahrungsstufen bei diagnostischen Aufgaben und bei der Planung des klinischen Managements, von publizierten Fallvignetten bis zur Beurteilung neuer Patientinnen und Patienten in der Notaufnahme. Das Modell übertraf über alle Szenarien hinweg sowohl die ärztlichen Vergleichsgruppen als auch ältere KI-Systeme (Brodeur et al. 2026). Die Autoren sehen darin einen Hinweis, dass LLMs die meisten etablierten Maßstäbe der klinischen Entscheidungsfindung inzwischen erreichen oder übertreffen, und fordern dringend prospektive Studien.

Diese Ergebnisse sind sowohl wissenschaftlich als auch medial aufmerksam verfolgt worden und beflügeln die Vision des langfristigen Potenzials solcher Systeme. In der klinischen Praxis entstehen allerdings häufig Herausforderungen durch vage Symptomschilderungen, grenzwertige diagnostische Ergebnisse oder atypische Befundkonstellationen. Diagnostische Leistungsfähigkeit ist allein kein hinreichendes Kriterium für den klinischen Einsatz: Fragen der Nachvollziehbarkeit, der Kontextsensitivität und der Integration in die ärztliche Verantwortung bleiben entscheidend und müssen den Einsatz solcher Systeme kontinuierlich begleiten.

Über die punktuelle diagnostische Treffsicherheit hinaus zeigen neuere Arbeiten, dass transformerbasierte (auf „selbstaufmerksamen" neuronalen Netzwerkarchitekturen zur Verarbeitung sequenzieller Daten beruhende) Modelle in der Lage sind, Krankheitsverläufe über längere Zeiträume hinweg abzubilden und zeitliche Abhängigkeiten zwischen Erkrankungen zu modellieren (Shmatko et al. 2025). Auf der Grundlage großer populationsbasierter Gesundheitsdaten können so Muster von Multimorbidität, Krankheitsprogression und konkurrierenden Risiken sichtbar gemacht werden. Einzelne Modelle können die Inzidenz und Entwicklung von über 1000 Erkrankungen vorhersagen, mit einer Vorhersagegüte, die der etablierter krankheitsspezifischer Modelle vergleichbar ist. Solche Ansätze eröffnen neue Perspektiven für Risikostratifizierung und Prognoseabschätzung, machen jedoch zugleich deutlich, dass mit zunehmender Modellkomplexität Fragen der Erklärbarkeit, der Übertragbarkeit auf individuelle Patientinnen und Patienten sowie möglicher Verzerrungen durch Trainingsdaten an Gewicht gewinnen.

Neben diesen inhaltlichen Überlegungen ist die technische und organisatorische Einbettung in den klinischen Alltag eine wesentliche Voraussetzung für einen verlässlichen Einsatz. Erst die Integration in Patientendatenmanagementsysteme (PDMS) und Krankenhausinformationssysteme (KIS) ermöglicht eine kontextgebundene, in bestehende klinische Abläufe eingebettete Nutzung. Zugleich erhöht eine solche Integration den Implementierungsaufwand und stellt neue Anforderungen an Anwendende, IT-Infrastruktur und institutionelle Steuerung. Der potenzielle Qualitätsgewinn ist stets gegen diesen Aufwand abzuwägen. Erste Implementierungen in großen klinischen Informationssystemen zeigen, dass eine solche Integration technisch realisierbar ist und in Pilotprojekten bereits erprobt wird. Eine systematische Evaluation dieser Ansätze hinsichtlich klinischer Wirksamkeit, Nutzerakzeptanz und Patientensicherheit steht jedoch weitgehend aus.

2.3 Verbesserte Patientenkommunikation

Ein weiteres relevantes Einsatzfeld generativer KI liegt in der Unterstützung der ärztlichen Kommunikation mit Patientinnen und Patienten. Generative KI ist in der Lage, komplexe medizinische Sachverhalte aus der fachlichen Terminologie in ein laienverständliches, strukturiertes und sprachlich angepasstes Format zu übertragen. Dies betrifft insbesondere Befunde, Diagnosen und Therapieoptionen, deren Bedeutung für Patientinnen und Patienten im klinischen Alltag häufig nur eingeschränkt nachvollziehbar ist. Dabei muss sichergestellt werden, dass die sprachliche Vereinfachung nicht zu einem Verlust an medizinischer Präzision führt.

Durch die Vereinfachung und Kontextualisierung medizinischer Informationen können LLMs dazu beitragen, die Gesundheitskompetenz von Patientinnen und Patienten zu stärken (Aydin et al. 2024). Verständlich formulierte Befundzusammenfassungen oder

erläuternde Begleittexte erleichtern das Verständnis der eigenen Erkrankung und fördern eine realistische Einschätzung von Risiken, Prognosen und Behandlungszielen (Kim et al. 2025). Für Internistinnen und Internisten eröffnet sich hier die Möglichkeit, die ärztliche Aufklärung gezielt zu ergänzen, ohne sie zu ersetzen. Voraussetzung hierfür ist eine validierte Datenbasis, um das Risiko von KI-generierten Fehlinformationen (sog. „Halluzinationen") zu minimieren. Allerdings zeigen empirische Untersuchungen mit Patientinnen und Patienten, dass diese insbesondere im Kontext akuter Erkrankungen vorrangig den direkten menschlichen Kontakt suchen und digitale Systeme nur als Ergänzung akzeptieren (Klein et al. 2026).

Darüber hinaus kann generative KI bei der Erstellung von Aufklärungsunterlagen, Merkblättern und individualisierten Informationsmaterialien unterstützen sowie häufige Patientenfragen systematisch aufgreifen und sowohl personalisiert als auch verständlich beantworten. Insbesondere in zeitlich stark verdichteten Versorgungssituationen kann dies helfen, Informationslücken zu schließen und die Kontinuität der Kommunikation zu verbessern – vorausgesetzt, die Inhalte werden fachlich geprüft und an den individuellen Kontext angepasst (Moëll und Sand Aronsson 2025).

Nicht zuletzt können verständliche und personalisierte KI-generierte Erläuterungen die gemeinsame Entscheidungsfindung unterstützen. *Shared Decision-Making* setzt voraus, dass Patientinnen und Patienten Optionen, Nutzen und Risiken nachvollziehen können. Generative KI-Systeme können dazu beitragen, diese Voraussetzungen zu schaffen, indem sie komplexe Entscheidungsalternativen strukturiert aufbereiten und transparent sowie vergleichbar darstellen. Dieser Prozess muss jedoch neutral gestaltet sein, um eine manipulative Beeinflussung (sog. „Nudging") durch die KI-Architektur auszuschließen. Die KI agiert hierbei rein informierend, während die abschließende Bewertung stets in das ärztliche Gespräch eingebettet und unter Wahrung der ärztlichen Letztverantwortung erfolgt. Studien müssen hier auch zeigen, dass KI-generierte Informationen korrekt verstanden werden und digitale Angebote patientenseitig tatsächlich genutzt werden. Dafür ist es, wie auch in anderen Zusammenhängen, notwendig, geeignete Routinedaten verfügbar zu machen, die ein kontinuierliches Monitoring solcher Prozesse erlauben.

Dabei ist zu differenzieren: Die verständliche Aufbereitung bereits ärztlich geprüfter Informationen ist anders zu bewerten als die dialogische Beantwortung individueller Patientenfragen oder die Unterstützung komplexer Aufklärungs- und Entscheidungsprozesse. Die situationsangemessene Gesprächsführung, die emotionale Einordnung, das Eingehen auf Rückfragen sowie die Berücksichtigung individueller Präferenzen und Belastungen bleiben genuin ärztliche Aufgaben; KI-gestützte Kommunikation kann das ärztliche Gespräch ergänzen, aber weder die persönliche Aufklärung noch die gemeinsame Entscheidungsfindung ersetzen.

2.4 Zeitersparnis und Qualitätssicherung durch KI-gestützte Ersteinschätzung in der Notaufnahme

Ein klinisch relevantes Einsatzfeld von generativer KI liegt in der Unterstützung von Ersteinschätzungs-, Priorisierungs- und Entscheidungsprozessen in der Notaufnahme (Williams et al. 2025). Gerade hier treffen eine hohe Arbeitsdichte, erheblicher Zeitdruck sowie eine große Bandbreite an Krankheitsbildern mit potenziell stark variierendem Schweregrad aufeinander.

LLMs könnten auf Basis strukturierter Eingangsdaten wie etwa Leitsymptomen, Vitalparametern, Vorbefunden oder Medikationslisten dazu beitragen, erste diagnostische Hypothesen zu strukturieren, relevante Differenzialdiagnosen sichtbar zu machen und auf potenziell kritische Konstellationen hinzuweisen. Dabei agiert die KI primär als kognitive Unterstützung und nicht als eigenständige Instanz. Entscheidend ist außerdem die zuverlässige Erhebung von Beschwerden und Symptomen. Sprachbarrieren, kognitive Beeinträchtigungen, die ggf. im Kontext des Notfalls auftreten, oder kulturelle Variationen der Beschwerdeäußerung können zu kritischen Fehleinschätzungen führen. Erfolgreich eingesetzt, könnten KI-gestützte Systeme sowohl Zeitgewinne bringen als auch zur diagnostischen Qualitätssicherung beitragen, etwa indem sie atypische Präsentationen häufiger Erkrankungen oder seltene, aber relevante Differenzialdiagnosen frühzeitig sichtbar machen. Insbesondere in Situationen mit hohem Patientenaufkommen und begrenzten personellen Ressourcen könnte dies das ärztliche Team entlasten und die Priorisierung klinischer Maßnahmen unterstützen (Vrdoljak et al. 2025). Voraussetzung hierfür ist jedoch die nahtlose Integration in bestehende klinische Workflows und etablierte Triage-Standards.

Unterstützung durch KI kann in der Notaufnahme sehr unterschiedliche Aufgaben umfassen. Dabei muss strikt zwischen rein informativer diagnostischer Strukturierung einerseits und formaler Ersteinschätzung (Triage), Priorisierung klinischer Maßnahmen und unmittelbarer Handlungssteuerung andererseits unterschieden werden. Je stärker ein System in diese Prozesse eingreift, desto höher sind die Anforderungen an Validierung, Transparenz, lokale Evaluation und ärztliche Kontrolle.

Solche Systeme sollten daher ausschließlich nach Validierung und lokaler Evaluation in regulierten Settings unterstützend zum Einsatz kommen, wobei die Verantwortung für Ersteinschätzung, Priorisierung und klinische Entscheidungen jederzeit bei qualifizierten medizinischen Fachpersonen verbleiben muss. Zudem ist die Evidenzlage bezüglich eines klinischen Nutzens sowie der sinnvollen Implementierbarkeit in den Versorgungskontext derzeit noch limitiert. Aktuelle Studien zeigen, dass z. B. die Nutzung von Symptom-Checkern im Wartebereich der Notaufnahme (Schmieding et al. 2025) die nachfolgende Versorgung nicht verbessert und dass die digitale Ersteinschätzung eher mehr Fälle in die Notaufnahme steuern und gleichzeitig kritische Konstellationen übersehen kann (Slagman et al. 2024). Darüber hinaus erwarten Patientinnen und Patienten klar eine primäre menschliche Interaktion im Notfallkontext (Klein et al. 2026).

3. Herausforderungen und Risiken für den Einsatz von LLMs im klinischen Kontext

3.1 Datenschutz und Datensicherheit

Der Einsatz generativer KI im klinischen Kontext berührt unmittelbar den Schutz besonders sensibler Gesundheitsdaten und damit einen zentralen Bereich ärztlicher Verantwortung. Patientendaten

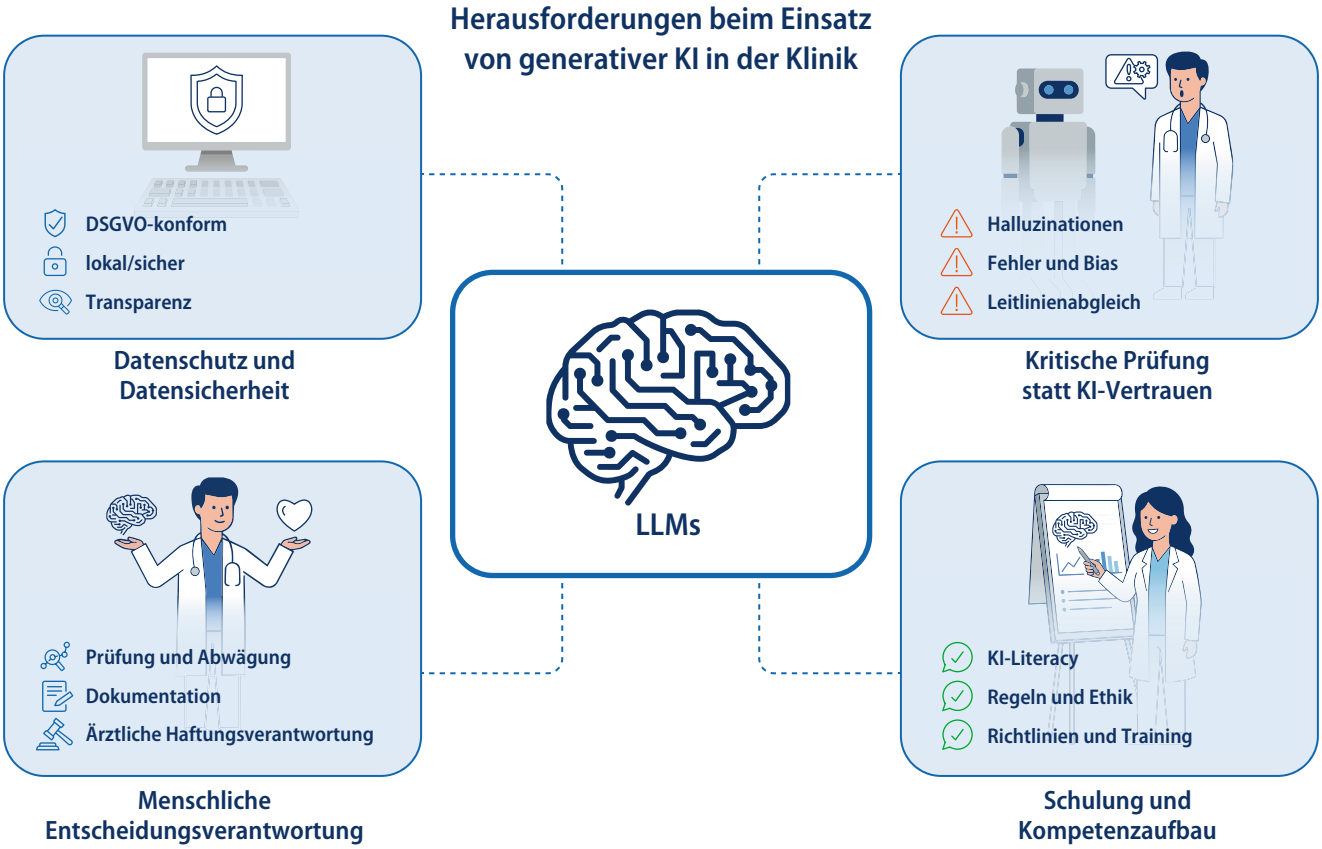


Abb. 2 ▲ Herausforderungen sowie datenschutzrechtliche und organisatorische Rahmenbedingungen für den Einsatz von generativer KI im klinischen Kontext. Die Abbildung verdeutlicht, dass Datenschutz, Nachvollziehbarkeit und klare Verantwortungsstrukturen beim Einsatz KI-gestützter Systeme untrennbar mit der ärztlichen Sorgfaltspflicht verbunden sind. Zentrale Voraussetzung für einen verantwortungsvollen Einsatz sind fundierte Schulungen und ein systematischer Aufbau von KI-Kompetenz (KI Literacy), der bereits früh im Medizinstudium beginnen sollte

unterliegen besonders hohen rechtlichen und ethischen Anforderungen; ihre Verarbeitung muss jederzeit zweckgebunden, nachvollziehbar sowie technisch und organisatorisch angemessen abgesichert erfolgen (Luckner und Lauer 2025). Vor diesem Hintergrund ist es grundsätzlich unzulässig und fachlich nicht vertretbar, personenbezogene oder anderweitig identifizierende Informationen in frei zugängliche oder datenschutzrechtlich ungeprüfte cloudbasierte Modelle einzugeben (Ong et al. 2024).

Für den klinischen Einsatz sind daher datenschutzkonforme und informationssichere Systemarchitekturen zu bevorzugen (Freyer et al. 2024). Dazu zählen insbesondere lokal betriebene Lösungen oder vertraglich und technisch abgesicherte Umgebungen, die den Anforderungen der DSGVO sowie des nationalen Datenschutz- und des ärztlichen Berufsrechts entsprechen. Je nach Anwendungsszenario können hierbei auch Open-Source-basierte Modelle in kontrollierten Infrastrukturen eine sinnvolle Option darstellen, sofern Informationssicherheit, Zugriffskontrolle, Protokollierung sowie ein klar definiertes Update- und Wartungsmanagement professionell umgesetzt sind (s. ◘ **Abb. 2**).

Dabei ist zwischen unterschiedlichen Nutzungsszenarien klar zu unterscheiden. Die Verwendung von generativer KI für allgemeine Wissensabfragen, Formulierungshilfen oder die Aufbereitung medizinischer Inhalte ohne Personenbezug ist datenschutzrechtlich anders zu bewerten als der Einsatz mit fallbezogenen, potenziell identifizierbaren oder unmittelbar personenbezogenen Gesundheitsdaten. Mit zunehmender Sensibilität der verarbeiteten Daten steigen die Anforderungen an Datenminimierung und Zugriffsschutz sowie an klare Freigabeprozesse und organisatorische Verantwortlichkeiten innerhalb der Einrichtung.

Unverzichtbar ist zudem ein hohes Maß an Transparenz gegenüber den Anwendenden. Diese müssen verlässlich darüber informiert sein, welche Daten verarbeitet werden, wo (technisch und geografisch) diese Verarbeitung stattfindet, wie lange Daten gespeichert werden und ob bzw. in welcher Form Eingaben zur Modellverbesserung genutzt werden. Eine solche Transparenz und Kenntnis der Rahmenbedingungen sind Voraussetzung für regelkonformes Handeln, für informierte Implementierungsentscheidungen der Einrichtungen und gegebenenfalls auch für eine angemessene Information der Patientinnen und Patienten (Huo et al. 2025).

3.2 Kritische Prüfung statt KI-Vertrauen

Wie eingangs dargelegt, liefert ein LLM keine aus einem Verstehen der Frage hervorgehenden Fakten, sondern die statistisch wahrscheinlichste Wortfolge. Daher sind LLMs trotz ihrer verführerisch hohen Sprachqualität und zunehmenden Leistungsfähigkeit „by design" und grundsätzlich fehleranfällig. Sie generieren Ausgaben auf Grundlage erlernter Wahrscheinlichkeitsverteilungen. Dadurch können sie sprachlich plausible, gleichwohl falsche Inhalte erzeu-

gen, ohne eigenes Nichtwissen oder Unsicherheit zuverlässig zu erkennen oder zu kommunizieren. Unzutreffende Schlussfolgerungen oder Inhalte sind daher auch häufig nicht auf den konkreten klinischen Kontext übertragbar, denn dieser fließt meist nicht voll umfänglich in die Frage ein. Diese systembedingte Fehleranfälligkeit erfordert ein besonders hohes Maß an kritischer Aufmerksamkeit und Validierung (Yu et al. 2024).

Jede durch ein LLM generierte Ausgabe ist grundsätzlich als Vorschlag oder Diskussionsgrundlage zu verstehen, nicht aber als verlässliche Handlungsempfehlung. Eine automatisierte Übernahme klinisch relevanter Inhalte ohne menschliche Kontrolle ist mit ärztlicher Sorgfalt, Verantwortung und Haftung nicht vereinbar. Generative KI kann vorarbeiten, strukturieren und diagnostische oder therapeutische Überlegungen anregen, darf aber nicht zur stillschweigenden Entscheidungsinstanz werden (Moura et al. 2024).

Für alle medizinisch relevanten Anwendungsfälle ist daher eine systematische und kontinuierliche Validierung der generierten Inhalte erforderlich. Diese müssen mit aktuellen Leitlinien, evidenzbasierten Quellen und etablierten klinischen Standards abgeglichen werden (Moëll und Sand Aronsson 2025). Dies gilt in besonderem Maße für diagnostische Überlegungen, Therapieempfehlungen und patientenbezogene Informationen. Nur durch eine solche kritische Prüfung und fachliche Einordnung lässt sich der Nutzen generativer KI erschließen, ohne die Qualität und Sicherheit der Patientenversorgung zu gefährden.

Kritische Prüfung darf dabei nicht allein als individuelle Pflicht der jeweils handelnden Ärztin oder des jeweils handelnden Arztes verstanden werden. Vielmehr setzt ein verantwortbarer Einsatz von LLMs auch institutionell definierte Einsatzgrenzen, klare Freigabe- und Prüfprozesse, angemessene Schulung sowie regulatorisch geforderte Formen menschlicher Aufsicht voraus. Nur wenn individuelle fachliche Validierung, institutionelle Governance und technische sowie rechtliche Rahmenbedingungen ineinandergreifen, lässt sich verhindern, dass KI-Vorschläge faktisch ungeprüft übernommen werden oder Verantwortlichkeiten unklar bleiben, was unmittelbare Folgen für die Versorgungsqualität hätte.

3.3 Menschliche Entscheidungsverantwortung

Ärztliches Handeln ist kein vollständig sequenzierbarer Vorgang, der maschinell einfach imitierbar wäre, sondern entsteht auch in der Interaktion mit Patientinnen und Patienten. Dabei sind Ärztinnen und Ärzte selbst im Sinne von Michael Balint therapeutisch wirksam (Balint 1965).

Voraussetzung für den ergänzenden und sinnvollen Einsatz generativer KI im klinischen Alltag ist daher stets eine klare konzeptionelle und praktische Trennung zwischen KI-generiertem Vorschlag und ärztlicher Entscheidung. Diese Trennung muss sowohl im individuellen Handeln als auch in den eingesetzten technischen Systemen, Benutzeroberflächen und klinischen Workflows nachvollziehbar abgebildet sein (Khera et al. 2024). Insbesondere darf nicht der Eindruck entstehen, dass klinische Entscheidungen automatisiert oder an Maschinen delegiert werden. Vielmehr ist sicherzustellen, dass KI-Outputs bewusst geprüft, fachlich eingeordnet und, sofern sie akzeptiert und übernommen werden, aktiv ärztlich verantwortet werden.

Auch aus rechtlicher Sicht ist die Verantwortungszuordnung eindeutig. Die Haftung für medizinische Entscheidungen verbleibt bei der ärztlich handelnden Person. Der Einsatz eines KI-Systems entbindet nicht von der Pflicht zur sorgfältigen fachlichen Validierung sowie zur nachvollziehbaren Dokumentation und Begründung von Entscheidungen (Angus et al. 2025). Daraus ergibt sich die Notwendigkeit, KI-Anwendungen so zu gestalten und einzusetzen, dass sie ärztliche Entscheidungsprozesse unterstützen, ohne die Verantwortlichkeit zu verschleiern oder faktisch zu relativieren (Palmieri et al. 2026).

Zugleich ist perspektivisch zu berücksichtigen, dass sich aus ärztlicher Sorgfaltspflicht nicht nur Grenzen, sondern auch potenzielle positive Handlungspflichten ergeben könnten (Angus et al. 2025). Wenn der Einsatz KI-gestützter Systeme in klar definierten Anwendungsfeldern einen nachweisbaren Nutzen zeigt, fachlich etabliert ist, von Kolleginnen und Kollegen zunehmend regelhaft eingesetzt wird und für die konkrete Versorgungssituation zumutbar erscheint, kann daraus eine Verpflichtung erwachsen, diese Möglichkeiten zumindest in die ärztliche Entscheidungsfindung einzubeziehen oder Patientinnen und Patienten über bestehende Alternativen zu informieren (Mello et al. 2025). Eine unterlassene Nutzung oder fehlende Aufklärung könnte in solchen Konstellationen künftig ebenfalls Gegenstand rechtlicher und ethischer Bewertung werden.

Je nach Anwendungskontext kann darüber hinaus eine transparente Kommunikation gegenüber Patientinnen und Patienten angezeigt sein, etwa wenn KI-Systeme in wesentlichem Umfang zur Vorbereitung von Dokumentation, Information oder Entscheidungsunterstützung beitragen (Mello et al. 2025). Eine solche Transparenz kann das Vertrauen stärken, setzt jedoch voraus, dass die Rolle der KI klar erläutert wird und die ärztliche Verantwortung für die Patientinnen und Patienten eindeutig erkennbar bleibt.

3.4 Schulung und Kompetenzaufbau

Ein sicherer und sinnvoller Einsatz generativer KI setzt voraus, dass Ärztinnen und Ärzte über eine angemessene KI-Kompetenz (KI-*Literacy*) verfügen. Diese geht über reine Anwenderkenntnisse hinaus und umfasst ein grundlegendes Verständnis der Funktionsweise, der Leistungsfähigkeit und der systemimmanenten Grenzen sowie typischer Fehlerquellen solcher Systeme (Cao et al. 2026). Entsprechend sollte KI-*Literacy* systematisch und strukturiert in Aus-, Weiter- und Fortbildungsstrukturen der Inneren Medizin integriert werden, sowohl in der klinischen Weiterbildung als auch in wissenschaftlichen Qualifikations- und Ausbildungswegen (Morgan et al. 2025).

Zentrale Inhalte solcher Qualifizierungsangebote sind unter anderem der sachgerechte Einsatz von *Prompting*-Techniken (d. h. der gezielten Formulierung von Eingaben zur Steuerung von KI-Modellen), die kritische Einschätzung der Qualität und Plausibilität von KI-*Outputs* sowie Kenntnisse zu rechtlichen, ethischen und datenschutzrechtlichen Rahmenbedingungen. Ziel ist es, Ärztinnen und Ärzte in die Lage zu versetzen, KI reflektiert und verantwortungsvoll zu nutzen, Risiken frühzeitig zu erkennen und die Systeme be-

wusst als unterstützende Instrumente einzusetzen (Schuitmaker et al. 2025). Dabei ist zu berücksichtigen, dass nicht alle Nutzergruppen dieselben Kompetenzen in gleicher Tiefe benötigen. Während für alle Ärztinnen und Ärzte eine Basiskompetenz im sachgerechten, kritischen und datenschutzkonformen Umgang mit KI erforderlich ist, bedürfen Personen in leitender, freigebender oder aufsichtsführender Funktion (z. B. Chef- und Oberärztinnen und -ärzte, IT-Leitungsverantwortliche) zusätzlicher Kenntnisse in Bewertung, Governance, regulatorischen Anforderungen und organisatorischer Implementierung.

Diese Anforderungen werden zunehmend auch regulatorisch konkretisiert. Nach Art. 4 der europäischen KI-Verordnung (EU AI Act) müssen Anbieter und Betreiber Maßnahmen zur Förderung der KI-Kompetenz der mit KI-Systemen befassten Personen treffen; dabei sind insbesondere technische Kenntnisse, Erfahrung, Aus- und Fortbildung sowie der jeweilige Nutzungskontext zu berücksichtigen. Ein bestimmtes individuelles Kompetenzniveau wird jedoch nicht verlangt. Ob ein medizinisches KI-System als Hochrisiko-System gilt, richtet sich nach Zweckbestimmung und regulatorischer Einordnung; der medizinische Einsatz allein genügt hierfür nicht. Für die Praxis bleiben technisches Grundverständnis, Risikobewusstsein sowie Kenntnisse der rechtlichen Rahmenbedingungen zentrale Bestandteile einer angemessenen Basiskompetenz (Schwartmann et al. 2025).

Auf institutioneller Ebene bedarf es darüber hinaus klarer organisatorischer Strukturen. Kliniken und größere Versorgungseinrichtungen sollten verbindliche interne Richtlinien zum Einsatz von KI-Anwendungen entwickeln, Verantwortlichkeiten definieren und geeignete Anlaufstellen etablieren, etwa in Form interdisziplinärer Arbeitsgruppen oder zentraler Beratungs- und Governance-Strukturen (Abdulnour et al. 2025). Ein solcher institutioneller Rahmen schafft Orientierung, fördert einen transparenten und konsistenten Umgang mit generativer KI und trägt dazu bei, individuelle Unsicherheiten zu reduzieren und die Patientensicherheit zu stärken.

Darüber hinaus empfehlen wir, bereits im Medizinstudium einen longitudinalen Pfad „KI in der Medizin" zu implementieren, der von den technischen Grundlagen über Anwendungen in Dokumentation und Wissensaufbereitung bis hin zur kritischen Reflexion von Entscheidungsunterstützung und agentischen Systemen reicht. So kann ein Kompetenzaufbau erfolgen, der nicht erst im späteren Berufsleben auf bereits etablierte Technologien reagieren muss, sondern zukünftige Ärztinnen und Ärzte frühzeitig zu einem souveränen und verantwortungsvollen Umgang mit solchen Systemen befähigt.

4. Prinzipien verantwortungsvoller Nutzung

4.1 Chancen nutzen trotz Restrisiko

Die klinische Realität zeigt, dass Produkte generativer KI bereits heute als Werkzeuge wahrgenommen und genutzt werden, um Arbeitsprozesse effizienter zu gestalten und ärztliche Tätigkeit stärker auf unmittelbar patientennahe Aufgaben auszurichten (Park et al. 2024). Angesichts der strukturellen Herausforderungen des Gesundheitssystems, wie steigender Komplexität, Ressourcenknappheit und wachsender Anforderungen an Qualität und Dokumentation, besteht ein legitimer Bedarf an digitalen Assistenzsystemen, die ärztliche Arbeit unterstützen, ohne sie zu substituieren (Thirunavukarasu et al. 2023).

Dabei ist jedoch zu berücksichtigen, dass unterschiedliche Einsatzfelder von LLMs auch unterschiedliche Risikoprofile aufweisen. Während risikoärmere, gut überprüfbare Anwendungen wie Strukturierungs- oder Formulierungshilfen vergleichsweise niedrigschwellig integriert werden können, gelten für entscheidungsnahe, patientenbezogene oder datensensible Anwendungen deutlich strengere Anforderungen an Validierung, Kontrolle und Freigabe. Fehlen solche geregelten Zugänge, kann dies dazu führen, dass generative KI dennoch genutzt wird, jedoch informell, unkoordiniert und außerhalb klarer Regelwerke. Eine solche „Schattennutzung" entzieht sich sowohl institutioneller Kontrolle als auch systematischer Qualitätssicherung und erhöht potenziell die Risiken für Datenschutz und Patientensicherheit (Odone et al. 2026).

Vor diesem Hintergrund halten wir eine proaktive und gestaltende Herangehensweise für erforderlich. Ziel muss es sein, generative KI bewusst in klar definierte und regulierte Versorgungsstrukturen einzubetten, Einsatzfelder präzise zu bestimmen und verbindliche Standards zu etablieren (Labkoff et al. 2024). Nur durch einen offenen, regulierten und reflektierten Umgang lassen sich die Chancen dieser Technologie nutzen, während Restrisiken transparent benannt, fortlaufend überwacht und möglichst reduziert werden (Palmieri et al. 2026).

4.2 Risiken kontrollieren, Grenzen reflektieren und akzeptieren

Ein verantwortungsvoller Umgang mit generativer KI setzt voraus, sich von der Vorstellung einer fehlerfreien oder autonomen „künstlichen Intelligenz" zu lösen (Jain et al. 2025). Trotz beeindruckender Leistungsfähigkeit bleiben LLMs probabilistische, statistisch basierte Systeme, deren Ausgaben nicht auf Verständnis, sondern auf Wahrscheinlichkeitsberechnungen über sprachliche Muster beruhen. Ein unreflektiertes Vertrauen in ihre vermeintliche Objektivität oder Perfektion ist daher mit den Anforderungen evidenzbasierter und verantwortungsvoller klinischer Medizin nicht vereinbar. In diesem Kontext ist und bleibt auch eine offene, positive Fehlerkultur im Umgang mit menschlichen Irrtümern unverzichtbar. Weder die vermeintliche eigene Perfektion noch die eigenen Unzulänglichkeiten werden durch KI-Systeme abgelöst.

Der Einsatz solcher Systeme erfordert klar definierte Bedingungen: eindeutig festgelegte Anwendungsgrenzen, eine klar kommunizierte Rollenverteilung zwischen KI-System und ärztlicher Entscheidung sowie Nutzerinnen und Nutzern, die sich der Stärken wie auch der Schwächen der Technologie bewusst sind (Labkoff et al. 2024). Wachsamkeit, kritische Distanz und fachliche Einordnung müssen dauerhaft integraler Bestandteil jeder Nutzung bleiben (Ratwani et al. 2024).

Risiken lassen sich nicht vollständig eliminieren, wohl aber systematisch identifizieren, überwachen und begrenzen. Dazu bedarf es geeigneter Governance-Strukturen, kontinuierlicher Schulung und klarer institutioneller Kontroll- und Verantwortungsmecha-

nismen (Rozenblit et al. 2025). Regelmäßige Evaluation, strukturierte Feedbackschleifen und eine fortlaufende Anpassung der Einsatzkonzepte sind notwendig, um Fehlentwicklungen frühzeitig zu erkennen und gegenzusteuern (Angus et al. 2025). Die bewusste Akzeptanz verbleibender Risiken, bei gleichzeitiger aktiver Risikominderung, ist Teil einer reflektierten, professionellen Haltung gegenüber dem Einsatz von generativer KI in der klinischen Medizin.

4.3 Ethisch-regulatorischer Rahmen

Der Einsatz generativer KI im klinischen Kontext ist eingebettet in einen bestehenden berufsrechtlichen und ethischen Rahmen, der uneingeschränkt fortgilt. Ärztliche Tätigkeit ist an die Prinzipien der Sorgfalt, der Verantwortung und des Handelns im Interesse der Patientinnen und Patienten gebunden. Diese Verpflichtungen bestehen unabhängig davon, ob digitale Assistenzsysteme eingesetzt werden, und sie lassen sich weder delegieren noch relativieren (Bratan et al. 2024).

Hinzu kommt eine Besonderheit der Medizin, die über reine Rechtsbefolgung hinausgeht: Ärztinnen und Ärzte stehen in einer gesellschaftlich verankerten besonderen Vertrauensstellung. Diese gründet nicht allein in gesetzlichen und berufsrechtlichen Vorgaben, sondern auch in einer öffentlich wahrnehmbaren professionellen Selbstverpflichtung, nämlich dem Anspruch, Patientinnen und Patienten in ihrer Verletzlichkeit verlässlich zu schützen und Schaden abzuwenden (*primum non nocere*). Für den Einsatz von KI bedeutet dies, dass Systeme, die ärztliches Denken, Priorisieren und Entscheiden wesentlich beeinflussen, so entwickelt, bereitgestellt und überwacht werden müssen, dass sie diese Vertrauensgrundlage nicht unterminieren, sondern aktiv unterstützen und sichern. Daraus folgt eine besondere Verantwortung der Entwickler, Hersteller und Inverkehrbringer, einschließlich klarer Anforderungen an Transparenz, Sicherheit, Prüfprozesse und Rechenschaftspflichten.

In diesem Zusammenhang ist auch § 2 Abs. 4 der (Muster-)Berufsordnung der Bundesärztekammer zu berücksichtigen, wonach Ärztinnen und Ärzte hinsichtlich ihrer ärztlichen Entscheidungen keine Weisungen von Nichtärzten entgegennehmen dürfen. Dieses Weisungsverbot schließt die Nutzung beratender Systeme nicht aus, macht jedoch deutlich, dass auch leistungsfähige KI-Anwendungen keine normsetzende oder verbindliche Entscheidungsrolle einnehmen dürfen. Selbst dann, wenn KI-Systeme in Studien eine hohe oder in einzelnen Aspekten überlegene Leistungsfähigkeit zeigen, bleibt ihre Funktion die der unterstützenden Entscheidungsassistenz. Ärztliche Entscheidungen müssen autonom getroffen, begründet und verantwortet werden; eine faktische Delegation klinischer Entscheidungsverantwortung an algorithmische Empfehlungen wäre berufsrechtlich und ethisch nicht vertretbar.

Gleichzeitig verlangt das Berufsrecht nicht nur die Wahrung ärztlicher Autonomie, sondern auch die Orientierung am jeweils anerkannten Stand der medizinischen Erkenntnis und an hohen Qualitätsstandards der medizinischen Versorgung. Vor diesem Hintergrund stellt sich perspektivisch die Frage, ob und in welchen Konstellationen der Einsatz KI-gestützter Systeme selbst Bestandteil sorgfältigen und leitliniengerechten ärztlichen Handelns werden kann (Smith et al. 2024). Wie bereits dargelegt, könnte unter solchen Bedingungen auch eine pauschale Nichtnutzung etablierter und nachweislich nützlicher KI-Anwendungen begründungsbedürftig werden.

Absehbar ist zudem, dass der regulatorische Rahmen in den kommenden Jahren weiter konkretisiert und ausdifferenziert wird. Auf europäischer Ebene werden mit dem EU AI Act und begleitenden Regelwerken zunehmend spezifische Anforderungen an Entwicklung, Einsatz und Überwachung von KI-Systemen formuliert werden (Luckner und Lauer 2025). Diese Entwicklungen werden den klinischen Einsatz von generativer KI weiter strukturieren, ersetzen jedoch nicht die Notwendigkeit, bereits heute verantwortungsvolle und praxistaugliche Maßstäbe anzulegen und proaktiv umzusetzen.

Dies gilt in besonderem Maße für Systeme mit agentischen Eigenschaften. Wenn KI nicht nur Inhalte generiert, sondern Aufgaben über mehrere Schritte hinweg eigenständig plant, Teilaufgaben delegiert oder unter Nutzung externer Werkzeuge bearbeitet, steigt die Komplexität der Gesamtkette. Für die Medizin bedeutet dies insbesondere, dass Restrisiken dann nicht mehr nur einzelne Ausgaben betreffen, sondern auch die Verkettung von Zwischenschritten, die Nachvollziehbarkeit des Prozessverlaufs und die klare Zuordnung menschlicher Verantwortung.

Bis zu einer weitergehenden rechtlichen Ausdifferenzierung bietet sich daher ein pragmatischer, kontextsensitiver Grundsatz an: *So viel KI wie verantwortbar und sinnvoll, so viel menschliche Beteiligung wie für Qualität, Sicherheit und Beziehung erforderlich*. Dieser Leitsatz bringt zum Ausdruck, dass technologische Unterstützung insbesondere in standardisierbaren und administrativen Prozessen umfassend genutzt werden kann, während in der direkten Arzt-Patient-Interaktion sowie bei komplexen klinischen Entscheidungen eine substanzielle menschliche Einbindung unverzichtbar bleibt.

5. Fazit und Handlungsempfehlung

Große Sprachmodelle und verwandte generative KI eröffnen klare Möglichkeiten, ärztliche Arbeit im klinischen Alltag zu unterstützen und Versorgungsprozesse effizienter zu gestalten. Vor dem Hintergrund der aktuellen Rahmenbedingungen im Gesundheitswesen ist ein grundsätzlicher Verzicht auf solche Systeme weder zeitgemäß noch realistisch. Gleichzeitig ist ihr Einsatz kein Selbstzweck, sondern bedarf einer bewussten, fachlich fundierten und verantworteten Einordnung. Dabei ist zu berücksichtigen, dass Nutzen, Risiken und Anforderungen an Prüfung und Kontrolle je nach Anwendung erheblich variieren können.

Medizinische Aufgaben sollten grundsätzlich von denjenigen Akteuren übernommen werden, die sie unter gegebenen Bedingungen am besten erfüllen können, stets mit dem Ziel, die bestmögliche Versorgung der Patientinnen und Patienten zu gewährleisten. Die Bewertung, wer oder was eine Aufgabe „am besten" erfüllen kann, ist in zweifacher Hinsicht zu bestimmen, nämlich in einer medizinischen und in einer ethischen Dimension. Die medizinische Dimension lässt sich empirisch prüfen, etwa durch Studien zur Qualität, Sicherheit und Effizienz von Verfahren. Die ethische Dimension hingegen erfordert verbindliche normative Leitplanken, die sicherstellen, dass medizinisches Handeln – unabhängig da-

von, ob es durch Menschen oder technische Systeme unterstützt wird – den gesellschaftlichen und berufsethischen Anforderungen entspricht.

Produkte generativer KI sind in diesem Sinne als leistungsfähige, aber nicht autonome Werkzeuge zu verstehen, nicht mehr, aber auch nicht weniger. Ihr Nutzen entfaltet sich nur dann, wenn sie von Ärztinnen und Ärzten bewusst eingesetzt, kritisch geführt und verantwortet werden. Mit der zunehmenden Leistungsfähigkeit medizinischer Technologien wächst zugleich die Verpflichtung, Rollen und Verantwortlichkeiten eindeutig zu definieren: Klinische Entscheidungen müssen nachvollziehbar bleiben, und Verantwortung darf nicht diffundieren. Ein verantwortbarer Einsatz sollte daher risikoadaptiert erfolgen und sich an der Art der Anwendung, Sensibilität der verarbeiteten Daten sowie Nähe zu diagnostischen oder therapeutischen Entscheidungen orientieren.

Unverändert maßgeblich bleiben die ärztlichen Standards. Patientensicherheit, fachliche Sorgfalt, Datenschutz und persönliche Verantwortung bilden den verbindlichen Rahmen, innerhalb dessen generative KI eingesetzt werden kann. Da KI-Systeme selbst keine Träger von Verantwortung oder Haftung sind, kommt Entwicklern, Herstellern und Anbietern eine besondere Verantwortung zu, die durch Regulierung, Transparenzanforderungen und Qualitätssicherung verbindlich eingefordert werden muss. Darüber hinaus wird zunehmend betont, dass bereits im Entwicklungsprozess eine bewusste Ausrichtung (sog. *Alignment*) von KI-Systemen an grundlegenden menschlichen und gesellschaftlichen Werten erfolgen sollte. Hierzu zählen insbesondere Prinzipien wie Patientensicherheit, Schadensvermeidung, Fairness, Transparenz und Rechenschaftspflicht. Neben regulatorischen Vorgaben gewinnt daher auch die freiwillige Selbstverpflichtung von Entwicklern und Unternehmen an Bedeutung, ethische Leitlinien und verantwortungsvolle Entwicklungspraktiken systematisch in Design, Training und Evaluation von KI-Systemen zu integrieren.

Der Einsatz von KI in der Medizin muss daher nicht nur rechtlich zulässig, sondern auch im Kern vertrauenswürdig und verantwortbar sein, wobei diese Vertrauenswürdigkeit für alle Beteiligten verbindlich sicherzustellen ist. Das gilt besonders dort, wo Systeme ärztliche Entscheidungen maßgeblich vorbereiten oder beeinflussen.

Die DGIM spricht sich daher für eine informierte, reflektierte und verantwortungsvolle Integration von generativer KI in die klinische Praxis aus. Voraussetzung sind klar definierte Anwendungsfelder, datenschutzkonforme technische Lösungen, die konsequente ärztliche Prüfung aller KI-generierten Inhalte sowie tragfähige Strukturen für Schulung, Governance und Aufsicht. Unter diesen Bedingungen kann generative KI zukünftig einen substanziellen Beitrag zur Unterstützung klinischer Entscheidungs- und Arbeitsprozesse leisten.

Leitend bleibt dabei der Anspruch, technologische Potenziale gezielt zu nutzen, Risiken systematisch zu begrenzen und den Einsatz konsequent an den Bedürfnissen und Interessen der Patientinnen und Patienten auszurichten.

DGIM-Kernpositionen

Grundsätze für den verantwortungsvollen Einsatz von generativer KI in der Inneren Medizin

1. Generative KI kann die internistische Versorgung sinnvoll unterstützen, darf ärztliche Urteilskraft und Verantwortung jedoch nicht ersetzen. Die Entscheidung, wie KI eingesetzt werden soll, muss aktiv in der Gesellschaft gefällt werden (nicht alles, was möglich ist, darf auch sein). Die DGIM wird für die Innere Medizin ihre Empfehlungen weiterentwickeln.
2. Personenbeziehbare Patientendaten dürfen ausschließlich in datenschutzkonform freigegebenen und institutionell geprüften Systemen verarbeitet werden.
3. Jede medizinisch relevante KI-Ausgabe ist vor klinischer Nutzung verpflichtend ärztlich zu prüfen.
4. Hochrisikoanwendungen, insbesondere in Diagnostik, Therapieplanung und Aufklärung, sind nur in validierten, regulierten und institutionell freigegebenen Systemen zulässig.
5. Kliniken und Praxen müssen vor dem Einsatz von KI in der Patientenversorgung verbindliche Strukturen für Governance, Schulung und Monitoring etablieren.
6. Patientinnen und Patienten sind transparent zu informieren, wenn KI wesentlich zur Dokumentation, Information oder Entscheidungsunterstützung beiträgt.
7. Anbieter von KI-basierten Systemen müssen Transparenz über Funktionsweise, Protokollierung, Updates und Datenverwendung gewährleisten. Perspektivisch ist zu fordern, dass KI-Systeme ihre Empfehlungen nachvollziehbar begründen und die zugrunde liegenden Evidenzquellen überprüfbar angeben.
8. Der Einsatz von KI ist regelmäßig und systematisch auf Wirksamkeit, Sicherheit und Qualität zu überprüfen. Hierzu müssen auch geeignete Routinedaten nutzbar gemacht und ggf. fachspezifisch erzeugt werden.
9. Kompetenzen im kritischen Umgang mit KI sind verbindlich in die medizinische Ausbildung zu integrieren.
10. Die Generierung klinischer Evidenz zu generativen KI-Anwendungen ist als eigenständiges Forschungsfeld zu etablieren und methodisch weiterzuentwickeln.

Literatur

Abdulnour R-EE, Gin B, Boscardin CK (2025) Educational Strategies for Clinical Supervision of Artificial Intelligence Use. N Engl J Med 393(8):786–797. https://doi.org/10.1056/NEJMra2503232

Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B et al (2025) AI, Health, and Health Care Today and Tomorrow. JAMA 334(18):1650–1664. https://doi.org/10.1001/jama.2025.18490

Aydin S, Karabacak M, Vlachos V, Margetis K (2024) Large language models in patient education: a scoping review of applications in medicine. Front Med 11:1477898. https://doi.org/10.3389/fmed.2024.1477898

Balint M (1965) THE DOCTOR'S THERAPEUTIC FUNCTION. Lancet 1(7397):1177–1180. https://doi.org/10.1016/s0140-6736(65)92717-0

Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A et al (2025) Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA 333(4):319–328. https://doi.org/10.1001/jama.2024.21700

Bratan T, Schneider D, Funer F, Heyen NB, Klausen A, Liedtke W et al (2024) Unterstützung ärztlicher und pflegerischer Tätigkeit durch KI: Handlungsempfehlungen für eine verantwortbare Gestaltung und Nutzung. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 67(9):1039–1046. https://doi.org/10.1007/s00103-024-03918-1

Brodeur PG, Buckley TA, Kanjee Z, Goh E, Ling EB, Jain P et al (2026) Performance of a large language model on the reasoning tasks of a physician. Science 392(6797):524–527. https://doi.org/10.1126/science.adz4433

Cao W, Zhang Q, Liu J, Liu S (2026) From Agents to Governance: Essential AI Skills for Clinicians in the Large Language Model Era. J Med Internet Res 28:e86550. https://doi.org/10.2196/86550
Chase A, Most A, Sikora A, Smith SE, Devlin JW, Shaochen X et al (2026) Evaluation of large language models' ability to identify clinically relevant drug-drug interactions and generate high-quality clinical pharmacotherapy recommendations. Am J Health Syst Pharm 83(13):e494–500. https://doi.org/10.1093/ajhp/zxaf168
Chen SF, Alyakin A, Seas A, Yang E, Choi JJ, Lee JV et al (2026) LLM-assisted systematic review of large language models in clinical medicine. Nat Med. https://doi.org/10.1038/s41591-026-04229-5
Ertl G, Becker C, Franz O, Galle J, Menelaou O, Meyer I et al (2025) Die Innere Medizin 2025/2030. Innere Medizin 66:951–961. https://doi.org/10.1007/s00108-025-01975-5
Freyer O, Wiest IC, Kather JN, Gilbert S (2024) A future role for health applications of large language models depends on regulators enforcing safety standards. Lancet Digit Health 6(9):e662–e672. https://doi.org/10.1016/S2589-7500(24)00124-9
Hautz WE, Marcin T, Hautz SC, Schauber SK, Krummrey G, Müller M et al (2025) Diagnoses supported by a computerised diagnostic decision support system versus conventional diagnoses in emergency patients (DDX-BRO): a multicentre, multiple-period, double-blind, cluster-randomised, crossover superiority trial. Lancet Digit Health 7(2):e136–e144. https://doi.org/10.1016/S2589-7500(24)00250-4
Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T et al (2025) Large Language Models for Chatbot Health Advice Studies: A Systematic Review. JAMA Netw Open 8(2):e2457879. https://doi.org/10.1001/jamanetworkopen.2024.57879
Jain SS, Goto S, Hall JL, Khan SS, MacRae CA, Ofori C et al (2025) Pragmatic Approaches to the Evaluation and Monitoring of Artificial Intelligence in Health Care: A Science Advisory From the American Heart Association. Circulation 152(23):e433–e442. https://doi.org/10.1161/CIR.0000000000001400
Khera R, Oikonomou EK, Nadkarni GN, Morley JR, Wiens J, Butte AJ, Topol EJ (2024) Transforming Cardiovascular Care With Artificial Intelligence: From Discovery to Practice: JACC State-of-the-Art Review. J Am Coll Cardiol 84(1):97–114. https://doi.org/10.1016/j.jacc.2024.05.003
Kim S-I, Park J, Kim T, Seo W, Kim T, Cha WC, Hong H (2026) Enhancing patient participation in emergency department through patient-friendly clinical notes generated by large language models. Sci Rep 16(1):1409. https://doi.org/10.1038/s41598-025-31113-y
Klein J, Sörries P, Mutlugil Y, George C, Altendorf M, Möckel M, Müller-Birn C (2026) With, not For: Co-Designing a Patient-Facing AI Companion Concept for the Emergency Department Waiting Area. In: Oliver Shamma NDA, Candello H, Cesar P, Lopes P, Bozzon A et al (Hrsg) Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. CHI 2026: CHI Conference on Human Factors in Computing Systems. ACM, Barcelona Spain, S 1–27. https://doi.org/10.1145/3772318.3790638
Labinsky H, Nagler L-K, Krusche M, Griewing S, Aries P, Kroiß A et al (2024) Vignette-based comparative analysis of ChatGPT and specialist treatment decisions for rheumatic patients: results of the Rheum2Guide study. Rheumatol Int 44(10):2043–2053. https://doi.org/10.1007/s00296-024-05675-5
Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E et al (2024) Toward a responsible future: recommendations for AI-enabled clinical decision support. J Am Med Inform Assoc 31(11):2730–2739. https://doi.org/10.1093/jamia/ocae209
Lawrence K, Kuram VS, Levine DL, Sharif S, Polet C, Malhotra K, Owens K (2025) Informed Consent for Ambient Documentation Using Generative AI in Ambulatory Care. JAMA Netw Open 8(7):e2522400. https://doi.org/10.1001/jamanetworkopen.2025.22400
Luckner S, Lauer W (2025) Regulatorische Einordnung KI-basierter Produkte für die medizinische Anwendung auf Basis von EU AI Act und MDR/IVDR. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 68(8):854–861. https://doi.org/10.1007/s00103-025-04091-9
McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J et al (2025) Towards accurate differential diagnosis with large language models. Nature 642(8067):451–457. https://doi.org/10.1038/s41586-025-08869-4
Mello MM, Char D, Xu SH (2025) Ethical Obligations to Inform Patients About Use of AI Tools. JAMA 334(9):767–770. https://doi.org/10.1001/jama.2025.11417
Moëll B, Sand Aronsson F (2025) Harm Reduction Strategies for Thoughtful Use of Large Language Models in the Medical Domain: Perspectives for Patients and Clinicians. J Med Internet Res 27:e75849. https://doi.org/10.2196/75849
Mohajer-Bastami A, Moin S, Suhaib A, Ahmed AR, Pouwels S, Hajibandeh S et al (2025) Artificial intelligence in healthcare: applications, challenges, and future directions. A narrative review informed by international, multidisciplinary expertise. Front Digit Health 7:1644041. https://doi.org/10.3389/fdgth.2025.1644041
Morgan DJ, Rodman A, Goodman KE (2025) How Physicians Can Prepare for Generative AI. JAMA Intern Med 185(12):1407–1408. https://doi.org/10.1001/jamainternmed.2025.4914
Moura L, Jones DT, Sheikh IS, Murphy S, Kalfin M, Kummer BR et al (2024) Implications of Large Language Models for Quality and Efficiency of Neurologic Care: Emerging Issues in Neurology. Neurology 102(11):e209497. https://doi.org/10.1212/WNL.0000000000209497
Odone A, Barbati C, Amadasi S, Schultz T, Resnik DB (2026) Artificial intelligence and infectious diseases: an evidence-driven conceptual framework for research, public health, and clinical practice. Lancet Infect Dis 26(3):e152–e167. https://doi.org/10.1016/S1473-3099(25)00412-8
OECD (2026), The agentic AI landscape and its conceptual foundations, OECD Artificial Intelligence Papers 56. https://doi.org/10.1787/396cf758-en
Ong JCL, Chang SY, William W, Butte AJ, Shah NH, Chew LST et al (2024) Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health 6(6):e428–e432. https://doi.org/10.1016/S2589-7500(24)00061-X
Palmieri S, Robertson CT, Cohen IG (2026) New Guidance on Responsible Use of AI. JAMA 335(3):207–208. https://doi.org/10.1001/jama.2025.23059
Park Y-J, Pillai A, Deng J, Guo E, Gupta M, Paget M, Naugler C (2024) Assessing the research landscape and clinical utility of large language models: a scoping review. BMC Med Inform Decis Mak 24(1):72. https://doi.org/10.1186/s12911-024-02459-6
Ratwani RM, Bates DW, Classen DC (2024) Patient Safety and Artificial Intelligence in Clinical Care. JAMA Health Forum 5(2):e235514. https://doi.org/10.1001/jamahealthforum.2023.5514
Rozenblit L, Price A, Solomonides A, Joseph AL, Koski E, Srivastava G et al (2025) Toward responsible AI governance: Balancing multi-stakeholder perspectives on AI in healthcare. Int J Med Inform 203:106015. https://doi.org/10.1016/j.ijmedinf.2025.106015
Russ P, Bedenbender S, Einloft J, Meyer HL, Wenzel LT, Ganser A et al (2026) Potential of large language models for rapid clinical information support: evidence from acute kidney injury knowledge testing. Sci Rep 16(1):11224. https://doi.org/10.1038/s41598-026-46846-7
Schmieding ML, Kopka M, Bolanaki M, Napierala H, Altendorf MB, Kuschick D et al (2025) Impact of a Symptom Checker App on Patient-Physician Interaction Among Self-Referred Walk-In Patients in the Emergency Department: Multicenter, Parallel-Group, Randomized, Controlled Trial. J Med Internet Res 27:e64028. https://doi.org/10.2196/64028
Schuitmaker L, Drogt J, Benders M, Jongsma K (2025) Physicians' required competencies in AI-assisted clinical settings: a systematic review. Br Med Bull 153(1):ldae025. https://doi.org/10.1093/bmb/ldae025
Schwartmann R, Benedikt K, Reif Y et al (2025) Datenschutz im Internet: Rechtshandbuch. C.H.BECK. ISBN 978-3-406-78340-1
Shmatko A, Jung AW, Gaurav K, Brunak S, Mortensen LH, Birney E et al (2025) Learning the natural history of human disease with generative transformers. Nature 647(8088):248–256. https://doi.org/10.1038/s41586-025-09529-3
Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW et al (2023) Large language models encode clinical knowledge. Nature 620(7972):172–180. https://doi.org/10.1038/s41586-023-06291-2
Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M et al (2025) Toward expert-level medical question answering with large language models. Nat Med 31(3):943–950. https://doi.org/10.1038/s41591-024-03423-7
Slagman A, Bremicker A, Möckel M, Eienbröker L, Fischer-Rosinský A, Gries A (2024) Evaluation of an Automated Decision Aid for the Further Referral of Emergency Room Patients–A Prospective Cohort Study. Dtsch Ärztebl Int 121(21):703–709. https://doi.org/10.3238/arztebl.m2024.0191
Smith H, Downer J, Ives J (2024) Clinicians and AI use: where is the professional guidance? J Med Ethics 50(7):437–441. https://doi.org/10.1136/jme-2022-108831
Solmonovich RL, Kouba I, Lee JY, Demertzis K, Blitz MJ (2025) Physician awareness of, interest in, and current use of artificial intelligence large language model-based virtual assistants. PLoS One 20(5):e0320749. https://doi.org/10.1371/journal.pone.0320749
Song JW, Park J, Kim JH, You SC (2025) Large Language Model Assistant for Emergency Department Discharge Documentation. JAMA Netw Open 8(10):e2538427. https://doi.org/10.1001/jamanetworkopen.2025.38427
Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. Nat Med 29(8):1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Tripathi S, Sukumaran R, Cook TS (2024) Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. J Am Med Inform Assoc 31(6):1436–1440. https://doi.org/10.1093/jamia/ocad258

Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R et al (2025) Towards conversational diagnostic artificial intelligence. Nature 642(8067):442–450. https://doi.org/10.1038/s41586-025-08866-7

Vrdoljak J, Boban Z, Males I, Skrabic R, Kumric M, Ottosen A et al (2025) Evaluating large language and large reasoning models as decision support tools in emergency internal medicine. Comput Biol Med 192:110351. https://doi.org/10.1016/j.compbiomed.2025.110351

Williams CYK, Subramanian CR, Ali SS, Apolinario M, Askin E, Barish P et al (2025) Physician- and Large Language Model-Generated Hospital Discharge Summaries. JAMA Intern Med 185(7):818–825. https://doi.org/10.1001/jamainternmed.2025.0821

Yu K-H, Healey E, Leong T-Y, Kohane IS, Manrai AK (2024) Medical Artificial Intelligence and Human Values. N Engl J Med 390(20):1895–1904. https://doi.org/10.1056/NEJMra2214183