Large Language Modelle zur schnellen Vereinfachung der Eingabe von Qualitätssicherungsdaten: Performance-Test mit Echtdaten am Beispiel der Tumordokumentation in der Urologie

Aktuelle Urologie

2024

10.1055/a-2281-8015

ISSN 0001-7868

Thieme

Large Language Modelle zur schnellen Vereinfachung der Eingabe von Qualitätssicherungsdaten: Performance-Test mit Echtdaten am Beispiel der Tumordokumentation in der Urologie

Large Language Models for Rapid Simplification of Quality Assurance Data Input: Field Trial with Real Data in the Context of Tumour Documentation in Urology

Autorinnen/Autoren
Johannes Frank[1], Axel S. Merseburger[1], Johannes Landmesser[1], Silvia Brozat-Essen[1], Peter Schramm[2], Laura Freimann[3], Alexander Kleehaus[4], Christian Elsner[5]

Institute
1 Urology Department, University of Luebeck, Luebeck, Germany
2 Department for Neuroradiology, University of Luebeck, Luebeck, Germany
3 Data Protection, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany
4 IT Consulting, PricewaterhouseCoopers Switzerland, Zuerich, Switzerland
5 Center for Artificial Intelligence, University of Luebeck, Luebeck, Germany

Schlüsselwörter
Tumor Dokumentation, künstliche Intelligenz, KI, ChatGPT, Ökonomie

Key words
ChatGPT, artificial intelligence, AI, economics, oncology documentation

eingereicht 24.1.2024
akzeptiert nach Revision 29.2.2024
online publiziert 2024

Bibliografie
Akt Urol
DOI 10.1055/a-2281-8015
ISSN 0001-7868

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Korrespondenzadresse
Dr. Christian Elsner, MD, University of Luebeck, Center for Artificial Intelligence, Ratzeburger Allee 160, 23562 Luebeck, Germany
christian.elsner@uni-luebeck.de

ZUSAMMENFASSUNG

Einleitung Large Language Modelle (LLMs) wie ChatGPT haben innerhalb kürzester Zeit die Anwendung von künstlicher Intelligenz in die breite Anwendung gebracht. Neben vielen verschiedenen Use-Cases der Textgenerierung und Verarbeitung ist eine Anwendung die Extraktion von Daten aus vorhandenen Dokumenten und Gesprächen zur vereinfachten und automatisierten Befüllung von Formularen.

Zielsetzung Gerade im Bereich der Qualitätssicherung und Dokumentation von Tumorerkrankungen fällt aktuell ein hoher Arbeitsaufwand an, Daten unter verschiedenen Aspekten in leicht variierenden Formaten und unter Anwendung von Interpretationen wie z. B. der TNM-Klassifikation von Tumoren zu übertragen. Zur Beurteilung der Anwendbarkeit von LLMs unterstützen Prozessen in diesem Bereich fehlen jedoch Feldversuche mit Echtdaten, die eine Beurteilung der Effizienz und Praktikabilität ermöglichen. Diese Arbeit soll einen Performance-Test dazu umsetzen und beurteilen.

Methodik Es wurde ein Performance-Test mit N = 153 datenschutztechnisch und durch eine Ethikkommission zu dem Zweck freigegebenen Arztbriefen von 25 Patienten vorgenommen. Mit der öffentlich verfügbaren Version von ChatGPT 4.0 wurden dazu mit einem automatisierten Programmskript die Aufgaben der Extraktion eines Erstdiagnosedatums sowie gängiger Tumorklassifikationen vorgenommen. Die Ergebnisse wurden dann einzeln auf Richtigkeit geprüft. Daran wurde dann der Nutzen eines Systems zum geführten Support bei Aufgaben im Kontext der Tumordokumentation indikativ beurteilt. Weiterhin wurde das Vorgehen auch im Kontext von Betriebskosten sowie potenzieller Hürden bis zur Anwendbarkeit beurteilt.

Ergebnisse In Summe kommt die Arbeit zum Schluss, dass der Einsatz generativer KI in diesem Feld vielversprechend ist und bereits im untrainierten Zustand als Hilfe tauglich ist. In einer simplifizierten Kalkulation stehen Kosten von 35 Cent einer Wertschöpfung von 61,54 Euro gegenüber. Es wird jedoch auch klar, dass die KI nur unterstützend tätig sein kann und die richtige Einbettung mit vorgefertigten

spezifischen natürlichsprachigen Abfragen (= Prompts) und Werkzeugen in den Arbeitsablauf entscheidend für die Performance ist.

Schlussfolgerung Der Einsatz von generativer KI im Kontext von Such-, Übertragungs- und Interpretationsarbeiten bei der Erstellung einer Tumordokumentation ist ein vielversprechender Ansatz. Die Umsetzung muss jedoch in praktischer Anwendung eng begleitet werden und das beste Zusammenspiel zwischen Mensch und Maschine weiter evaluiert und mit spezifischen Werkzeugen begleitet werden.

ABSTRACT

Introduction Large Language Models (LLMs) such as ChatGPT have rapidly brought the application of artificial intelligence into widespread use. Among many different use cases for text generation and processing, one application is the extraction of data from existing documents and conversations for simplified and automated form-filling.

Objective In the field of quality assurance and documentation of cancer diseases, there is currently a significant workload involved in transferring data under various aspects into slightly varying formats and applying interpretations such as the TNM classification of tumours. However, there is a lack of trials with real data to assess the applicability of LLM-supported processes in this area, which would enable an evaluation of efficiency and practicality. This study aims to implement and assess such a trial.

Methodology A trial was conducted with N = 153 privacy-compliant and ethics committee-cleared medical reports from 25 patients. Using the publicly available version of ChatGPT 4.0, an automated script was used to extract the date of initial diagnosis and common tumor classifications. The results were then individually checked for accuracy. Based on this, the utility of a simple system for guided support in tasks related to tumour documentation was assessed. Additionally, the approach was evaluated in terms of operational costs for the model and its applicability.

Results In summary, the study concludes that the use of generative AI in this field is promising and suitable as a tool even in an untrained state. In a simplified calculation, costs of 35 cents are offset by a value creation of 61,54 euros. However, it also becomes clear that AI can only act in a supportive role, and the correct integration with premade specific prompts and tools into the workflow is crucial for a relevant performance.

Conclusion The use of generative AI in the context of search, transfer, and interpretation tasks in the creation of tumor documentation is a promising approach. However, its implementation in practical applications must be closely monitored, and the optimal interaction between man and machine should continue to be evaluated and must be accompanied by tools and task-specific prompts.

Einleitung – Large Language Modelle und deren potenzieller Impact in der Medizin

Trotz der breiten Auswirkungen von künstlicher Intelligenz (KI) in verschiedenen Industrien ist ihre Anwendung in der klinischen Versorgung noch begrenzt. Das liegt insbesondere an Herausforderungen wie dem Mangel an strukturierten, maschinenlesbaren Daten und der teils mangelhaften Interoperabilität zwischen Gesundheits-IT-Systemen. Neuere, sogenannte Large Language Modelle (LLMs) [1] zeigen hier aber Fähigkeiten, die über traditionelle KI-Modelle hinausgehen. Sie sind in der Lage, auch ohne spezifisches Training Lösungen für medizinische Fragestellungen zu fertigen bzw. mit Medizindaten sprachliche Strukturierungen zu verrichten [2].

Man lässt durch LLMs dabei verschiedene Aufgaben verrichten, so etwa: aus Texten Informationen zusammenfassen, nformationen extrahieren, die Informationen klassifizieren, semantische Suchen auf den Texten durchführen, fehlende Informationen ergänzen oder auch direkt Fragen über den Text beantworten [3]. Mit diesen technischen Fähigkeiten sind teilweise bereits mit allgemein trainierten Modellen auch erstaunliche medizinische Arbeiten gelungen. So hat z. B. eine Gruppe von Wissenschaftlern mit einer frühen Version von ChatGPT das amerikanische medizinische Staatsexamen absolvieren können, wobei etwas über 60% korrekte Antworten für alle 3 Teile des Examens erreicht wurden [4]. Weiterhin gibt es verschiedene Tests und Anwendungsfelder, medizinische Texte über diese Modelle zu fertigen. Ein Beispiel wäre der Versuch, aus einfachen Daten einer Krankenakte oder Vorbefunden Arztbriefe als allgemeinverständliche, zusammengefasste Fließtexte zu fertigen, was ebenfalls mit guten Ergebnissen gelingt [5].

Speziell die Fähigkeiten von LLMs bzw. dem Programm ChatGPT des Unternehmens OpenAI, Informationen aus einem Text zu klassifizieren und inklusive semantischer Umformung auch Informationen zu extrahieren, könnte für die Aufgabe der Tumordokumentation sehr nützlich sein. In diesem Feld wird heute von Tumordokumentaren viel Aufwand für die Aufbereitung von Freitextinformationen in meist abweichende Formate verwendet. Der größtenteils „human“ geführte Prozess extrahiert aus Standarddaten, allgemeinen medizinischen Informationen und Texten über den Patienten dann übersichtliche Daten wie z. B. die Metastasierung des Tumors, das Erstdiagnosedatum und weitere Klassifikationsdaten.

Diese Studie zielt darauf ab, eine Evaluation der Leistung von ChatGPT in der aktuellen Version im Zugriff per Internet am 29. Oktober 2023 (GPT 4.0) zu geben und sein Potenzial für die Integration in den Prozess der Tumordokumentation abzuschätzen. Dazu wurden Echtdaten von 25 Patienten mit Prostatakarzinom in Form von 153 Arztbriefen datenschutztechnisch freigegeben und dann genutzt, um die am 29.10.2023 verfügbare Version von ChatGPT mit 2 typischen Aufgaben der Tumordokumentation zu beauftragen.

Grundsätzliche Herausforderungen beim Einsatz von LLM-Modellen und deren Management in der vorliegenden Studie

Die Umsetzung des Einsatzes von generativer KI im Kontext der Verarbeitung medizinischer Daten muss in praktischer Anwendung eng begleitet werden und das beste Zusammenspiel sowohl unter ergonomischen als auch inhaltlichen bzw. medikolegalen Aspekten ausführlich evaluiert werden. Aktuell gibt es noch wenige Arbeiten, die dies systematisch mit der manuellen Validierung von LLM-verarbeiteten Daten vornehmen wie z. B. Johnson et al. [6]. Die Herausforderungen beim Einsatz von LLM-Systemen sind dabei an sich immer ähnlich relevant: Datenschutz und die Endverantwortung für die prozessierten medizinischen Daten sowie die De-facto-Einbettung in den Arbeitsablauf unter Aspekten der Ergonomie für den Nutzer.

Im Punkt des Datenschutzes kann man ausführen, dass der Einsatz von LLMs im medizinischen Sektor wichtige Grundsatzfragen aufwirft. Die heute in Betrieb befindlichen Lösungen sind oft nicht nur in einer öffentlichen Cloud betrieben, sondern auch die (öffentliche) Weiterverarbeitung von Daten in den Modellen wird nicht eindeutig verhindert. Eine mögliche Lösung ist die Verwendung von Daten, die vollständig mit der Datenschutz-Grundverordnung (DSGVO) konform sind – also die Nutzung von anonymisierten Daten. Allerdings stellt die Anonymisierung in komplexen medizinischen Akten eine Herausforderung dar und ist in Reinform quasi unmöglich. Eine alternative Lösung könnte der Betrieb einer eigenen LLM-Instanz auf lokalen oder gesicherten Rechenzentren einer Klinik sein. Dies würde die Kontrolle über die Datenhaltung erhöhen und könnte das Risiko eines Datenschutzverstoßes minimieren. Allerdings ist der Betrieb lokaler LLM-Lösungen aktuell noch mit Einschränkungen in der Leistung verbunden.

Beide Ansätze (Anonymisierung und lokale Instanzen) finden aber grundsätzlich Unterstützung in der Literatur, die sich mit Datenschutzkonzepten im Gesundheitswesen auseinandersetzt, wie beispielsweise bei Rumbold und Pierscionek (2017) in ihrem Artikel über Datenschutz und Gesundheitsanwendungen [7], die eben diese beiden grundsätzlichen Herangehensweisen vorschlagen.

Management des Datenschutzes – im vorliegenden Fall wurden die Daten vollständig pseudonymisiert und doppelt geschützt

Um auf die Leistungsfähigkeit der aktuellen ChatGPT 4.0 Version zugreifen zu können, haben sich die Autoren in der vorliegenden Arbeit dazu entschieden, den Weg von DSGVO-konform freigegebenen Daten zu gehen. Diese wurden nach Freigabe durch den lokalen Datenschutz der Klinik einer ausgewählten Gruppe an Bearbeiter*innen zur Verfügung gestellt. Die Daten selbst waren dabei einerseits von seit 10 Jahren verstorbenen Patienten verwendet worden und neben jeder Entfernung von direkt personenbezogenen Daten wurde eine zusätzliche manuelle Bearbeitung vorgenommen, um auszuschließen, dass aus der Krankengeschichte personenidentifizierende Merkmale gezogen werden können. Damit lag ein doppelter Schutz der Daten vor, da auf die Patienten im eigentlichen Sinne kein Datenschutz anzuwenden war, dieser aber trotzdem nach den strengen Regeln angewandt wurde. Zusätzlich wurden die Daten nur für den Zeitraum der Auswertung zur Verfügung gestellt und danach wieder vernichtet. Natürlich wurde die Arbeit auch bei der Ethikkommission der Universität zu Lübeck vorgelegt und von dort in der vorliegenden Form für eine Promotion freigegeben.

Eine andere Möglichkeit zur in der Studie verwendeten Vorgehensweise wäre tatsächlich die Generierung von synthetischen Daten gewesen. Diese Vorgehensweise wird oftmals verwendet, um Systeme zu testen und einen ersten Eindruck von der Praktikabilität eines Ansatzes zu bekommen [8]. Die Verwendung wird aber eben auch teils kritisch und allenfalls als Vorstufe zu echten Feldtests gesehen, da die Daten eben nicht ein „Real-world"-Szenario mit allen Schwächen und Beschaffenheiten „echter" Daten aufweisen bzw. diese Daten immer eine Vorprägung durch die synthetischen Methoden der Erstellung haben [9].

Endverantwortung für die Überprüfung und Übertragung von Daten in das Tumorregister liegt weiter beim Menschen selbst

Ein weiterer kritischer Aspekt ist die Endverantwortung für die von LLMs verarbeiteten medizinischen Daten. Trotz der fortschrittlichen Fähigkeiten von KI-Systemen zur Datenanalyse und -verarbeitung bleibt die Verantwortung für die Richtigkeit und Angemessenheit der medizinischen Dokumentation beim medizinischen Fachpersonal. Hier bietet sich ein prozessuales Handling an, bei dem eine „Human"-Instanz – also ein qualifizierter Mediziner – die von LLMs erarbeiteten Daten überprüft und freigibt. Dieser Ansatz entspricht den Empfehlungen zur klinischen Entscheidungsunterstützung durch KI, wie sie von Topol (2019) in seinem Überblick über KI im Gesundheitswesen vorgeschlagen wurden [10].

Durch diesen integrierten Ansatz, bei dem die KI als Unterstützungstool und nicht als Ersatz für menschliche Expertise fungiert, kann sichergestellt werden, dass die medizinische Dokumentation sowohl effizient als auch medikolegal adäquat ist.

Das Feld der Tumordokumentation – Aufwände und Ansatzpunkte für Automatisierungen

Die Datenpflege in einem klinischen Krebsregister erfordert Aufwand, der von verschiedenen Faktoren abhängt. Dazu gehören etwa die Größe des Registers, die Anzahl der zu erfassenden Fälle sowie die Komplexität der Daten und der Technologie, die verwendet wird. Generell umfasst die Datenpflege in einem Krebsregister die Erfassung, Validierung, Aktualisierung und Analyse von Daten. Dazu gehört:

Datenerfassung: Das Sammeln von Daten aus verschiedenen Quellen wie Patientenakten, Laboren und anderen medizinischen Einrichtungen. Hier können KI-Systeme durch Vorausfüllen und Vorselektieren von Daten unterstützen, wenn nicht sogar Daten automatisch extrahieren. Dies ist ein Ansatzpunkt dieser Studie.

Qualitätssicherung: Die Überprüfung der Genauigkeit und Vollständigkeit der Daten. Hier könnten allgemeine Prüfverfah-

ren auf KI- oder LLM-Basis greifen. Die Studie betrachtet dies nicht.

Datenaktualisierung: Regelmäßige Aktualisierung der Patientendaten, um den Behandlungsverlauf und den Gesundheitszustand der Patienten genau widerzuspiegeln. Soweit hiervon eine erneute Datenerfassung betroffen ist, deckt die Studie diesen Aspekt in Teilen ab.

Datenauswertung: Analyse der gesammelten Daten für Forschungszwecke und zur Verbesserung der Krebsbehandlung und -vorsorge. Hier könnten allgemeine Prüfverfahren auf KI- oder LLM-Basis z. B. bei der Zusammenfassung und Auswertung greifen. Die Studie betrachtet dies nicht.

Die Anzahl des benötigten Personals variiert je nach Umfang und Spezifität des Registers. In größeren Registern kann ein Team aus Datenmanagern, Statistikern, IT-Spezialisten und medizinischem Personal erforderlich sein, während kleinere Register möglicherweise mit weniger Personal auskommen. Die genannte Dokumentationsarbeit selbst beinhaltet oftmals die Befüllung verschiedener Daten- und Dokumentationsfelder, wobei vielfach Schemata der Tumorklassifikation, Datum der Erstdiagnose, Remissionsdatum und andere Daten extrahiert werden müssen. Die Tumordokumentation beinhaltet daher oftmals eine Art „Transposition" von Daten aus vorhandenen Angaben in andere Schemata. Für die Studie wurde dies auf 4 exemplarische Arbeitsbereiche angewendet, die in der ▶ **Tab. 1** dargestellt sind.

Materialien und Methoden bei der Überprüfung der ChatGPT-Funktion bei der Tumordokumentation

Setup eines Evaluationsansatzes für den Einsatz von ChatGPT im Kontext der Tumordokumentation

Es wurde ein Feldversuch mit N = 153 datenschutztechnisch freigegebenen Arztbriefen von 25 Patienten mit Prostatakarzinom vorgenommen. Mit der öffentlich verfügbaren Version von ChatGPT 4.0 wurden dazu mit einem automatisierten Programmskript die Aufgaben der Extraktion eines Erstdiagnosedatums sowie allgemeiner, tumorfokussierter Informationen (Anamnese, digital-rektaler Befund, prostataspezifisches Antigen [PSA], Bildgebungsbefund, Biopsiebefund) vorgenommen. Die Ergebnisse wurden dann durch medizinisches Fachpersonal einzeln auf Richtigkeit geprüft. Der Prozess wurde durch ein eigens entwickeltes Programm in Phython Script automatisiert und mit einer automatisierten Schnittstelle (API) von ChatGPT verarbeitet. Dabei wurden verschiedene Formulierungen der Anweisungen (Prompts) getestet, ein Prozess, der als Prompt Engineering bezeichnet wird. Die endgültige Fertigung der Zusammenfassungen erfolgte über einen Zugriff am 29. Oktober 2023 mittels der per Internet verfügbaren Version ChatGPT 4.0. Für das Thema der Tumorklassifikation wurden die 4 verschiedenen Abfrageanweisungen an ChatGPT gestellt, die dem jeweiligen Bereich der Biopsie/Pathologie zuzuordnen sind. In den Feldern B1–B4 der ▶ **Tab. 2** befindet sich die jeweilige Beschreibung für ChatGPT, welcher Befund extrahiert und in das Formular der Tumordokumentation gegeben werden sollte.

Die Validierung wurde nach verschiedenen Kategorien der richtigen Extraktion und richtigen Einordnung vorgenommen

Für die Evaluation von Potenzialen und Limitationen bei der KI-gestützten Tumordokumentation wurde ein mehrteiliger Ansatz gewählt. Einerseits ging es um die Frage, wie zuverlässig ist ChatGPT darin, bestimmte Informationen aus Texten zu filtern. Andererseits bedurfte es einer Evaluation aus klinischer Sicht mit der Frage nach dem qualitativen Wert der Zusammenfassungen.

Für die Bewertung der Zuverlässigkeit ChatGPTs bei der Filterung von Texten wurde eine objektivierbare Information gewählt: das Datum der Erstdiagnose des Prostatakarzinoms. Die Fragestellung lautete hier: Wie oft erkennt ChatGPT aus einer

▶ **Tab. 1** Erläuterung der 4 spezifischen Tumordokumentationsarten, die durch das LLM-Tool extrahiert und bearbeitet wurden.

Dokumentationsart	Beschreibung	Details/Elemente
Allgemeinbefund	Allgemeine Beschreibung des Gesundheitszustands des Patienten, einschließlich des Vorhandenseins von Tumoren.	▪ Patientengeschichte ▪ Untersuchungsergebnisse ▪ Symptome
Gleason-Score	Bewertungssystem zur Bestimmung der Aggressivität von Prostatakrebs.	Bewertung basierend auf mikroskopischer Untersuchung des Prostatagewebes, es ergibt sich ein Score zwischen 2–10 (im klinischen Alltag meist nur 6–10 verwendet)
UICC-Klassifikation	Internationale Klassifikation von Krebsstadien, entwickelt von der Union for International Cancer Control.	Stadieneinteilung nach Größe und Ausbreitung des Tumors, das auch Lymphknoten und Fernmetastasen berücksichtigt
TNM-Klassifikation	Tumorklassifikationssystem, das die Größe und Ausbreitung des Primärtumors (T), das Vorhandensein von Krebszellen in den regionalen Lymphknoten (N) und das Vorhandensein von Fernmetastasen (M) beschreibt.	T-Kategorie: Größe/Ausbreitung des Primärtumors N-Kategorie: Beteiligung regionaler Lymphknoten M-Kategorie: Fernmetastasen

▶ **Tab. 2** Auszug aus einfachen „Prompt"-Anweisungen zur Extraktion der verschiedenen Dokumentationsarten.

	Beschreibung für die API-Abfrage von ChatGPT per Python Script
B1	Biopsiebefund – Allgemeinbefund
B2	Biopsiebefund – Gleason-Score. Beschreibt die histologische Morphologie des Drüsenmusters der Prostata. Gleason 1–5 beschreibt das Drüsenmuster einer Prostataprobe. Der Gleason Score 2–10 errechnet sich aus dem Vergleich zweier Probestücke und ihrem jeweiligen Drüsenmuster.*
B3	Biopsiebefund – UICC-Klassifikation Stadium I–IV. Orientiert sich an TNM-Klassifikation. Beispiel: T0–2a entspricht UICC Stadium I.
B4	Biopsiebefund – Einteilung nach TNM-Klassifikation. T wird eingeteilt in 0–4 und beschreibt die Größe des Tumors, meistens in cm. N wird eingeteilt in 0–1 und beschreibt, ob und wie stark Lymphknoten befallen sind. M wird eingeteilt in 0–1 und beschreibt, ob Fernmetastasen bestehen. Beispiel: T2N1M0 = tastbares Prostatakarzinom mit regionalem Lymphknotenbefall ohne Fernmetastasen.*

*Beschreibung wurde mit verschiedenen, teils unvollständigen und oberflächlichen Angaben zur korrekten Klassifikation verarbeitet. Ein Neudurchlauf mit korrigierten Prompts laut den Angaben nach UICC, 8. Aufl. 2017 ergab für das Ergebnis keine relevanten Unterschiede, d. h. die KI hat hier bereits teils aus vortrainierten Angaben gearbeitet.

Menge an Arztbriefen denjenigen, in dem die Prostatadiagnose erstmalig gestellt wurde? Zusätzlich wurde für die Extraktion von Biopsiebefunden eine quantitativ-qualitative Evaluation vorgenommen. Dafür wurden für jeden Arztbrief die von ChatGPT angegebenen Inhalte der Biopsiebefunde (Allgemeinbefund, Gleason-Score, UICC-Klassifikation, TNM-Klassifikation) mit dem Originaltext verglichen. Diese Inhalte wurden dann in Bezug auf inhaltliche Korrektheit (A) sowie zeitlich und inhaltlich korrekte Einordnung (B) bewertet. Unter Punkt A wurde bewertet, ob Angaben grundsätzlich richtig übernommen wurden oder nur teilweise erkannt wurden. Unter B wurde dann differenziert, ob die Information auch zeitlich und inhaltlich richtig zugeordnet wurde. Ein Fehler in der inhaltlichen Zuordnung wäre beispielsweise die Einordnung eines TNM-Befunds als Gleason-Score. Eine zeitliche Fehlzuordnung bestand zum Beispiel dann, wenn ein alter Befund nicht als solcher markiert bzw. als aktuell dargestellt wurde (▶ **Tab. 3**).

▶ **Tab. 3** Zweigliedriges Schema der Evaluation der Richtigkeit der LLM-basierten Extraktion.

A: grundsätzlich richtige Extraktion und/oder Transposition der Information	B: richtige Einordnung im inhaltlichen und/oder zeitlichen Kontext
1 = Information richtig und vollständig extrahiert bzw. auch transponiert	1 = Information zeitlich und inhaltlich korrekt zugeordnet
2 = Information teilweise richtig extrahiert bzw. auch transponiert, nicht vollständig	2 = Information in Teilen nicht eindeutig richtig zugeordnet, Expertenmeinung nötig
3 = Information falsch bzw. nicht extrahiert, inhaltliche Fehler	3 = Information inhaltlich oder zeitlich nicht richtig zugeordnet
4 = keine Bewertung möglich	4 = keine Bewertung möglich
5 = keine Angabe im Brief/korrekte Nichtangabe	5 = keine Angabe im Brief/korrekte Nichtangabe

Ergebnisse des Einsatzes im Kontext der Tumordokumentation mit N = 153 Arztbriefen

Bezüglich der Fähigkeiten ChatGPTs, bestimmte Informationen aus einer Zahl an Dokumenten zu finden, lässt sich ein positives Urteil ziehen.

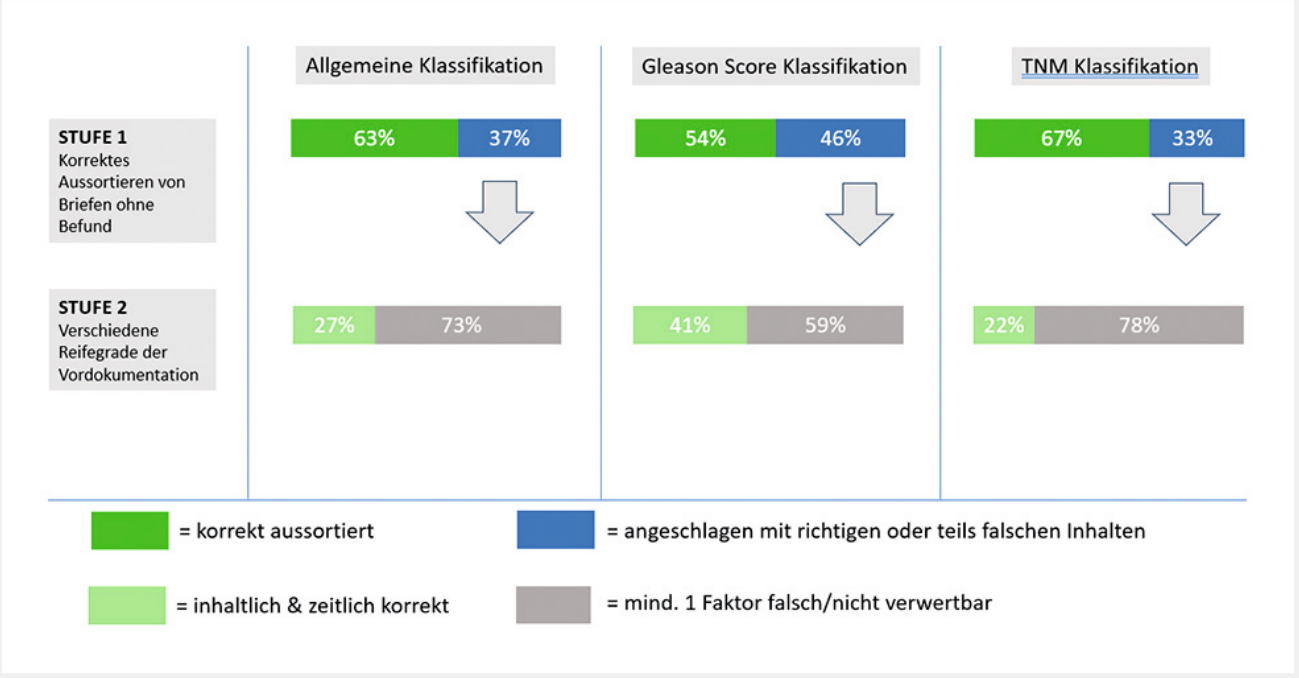
In 23 von 25 Fällen (92%) erkannte die KI korrekterweise das Entlassdatum des Arztbriefs, in dem die Diagnose Prostatakarzinom erstmals aufgeführt wurde. In 2 Fällen wurde das Datum nicht korrekt erkannt. Bei einem davon wurde in den Arztbriefen nur eine Prostatektomie ohne Karzinombefund beschrieben.

Die Auswertung der Extraktion von Biopsiebefunden wird in unserem Artikel übersichtsartig dargestellt. Es wurden die 4 Kategorien abgefragt, die in der ▶ **Tab. 1** bereits erläutert worden sind. In der Stichprobe war dabei auch eine größere Menge von Briefen, in denen keine Informationen zu den jeweiligen Kategorien aufgeführt vorlagen. Auch diese Selektion war bereits eine Supportaufgabe durch das LLM-Modell, mit der der Arbeitsaufwand gesenkt wird.

Das Modell schöpfte in dieser ersten Stufe dadurch Wert, dass in der Tat 100% aller als „ohne Befund" erkannten Briefe auch ohne Befund waren. Für die Allgemeinbefunde wurden damit 63%, für den Gleason-Score 46%, für die UICC-Klassifikation 97% und für die TNM-Klassifikation 67% der Arztbriefe korrekt und fehlerfrei aussortiert. Diese Briefe wurden für die nachstehende Darstellung in der ersten Stufe ausgewiesen.

In der zweiten Stufe wurden dann die vom LLM identifizierten Briefe, in denen auch tatsächlich Befunde erkannt werden konnten, bewertet: In der Kategorie Allgemeinbefunde wurden in den Briefen mit Befund durch das LLM zu 43% inhaltlich korrekte Angaben gemacht. Zu 21% waren die Ergebnisse teilweise korrekt und zu 34% waren sie falsch oder wurden nicht richtig erkannt. Bei ebenfalls 43% fand eine zeitlich und inhaltlich korrekte Einordnung statt. Teilweise korrekt war sie in 9% der Fälle und zu 23% war sie inkorrekt.

Bei der Extraktion des Gleason-Score waren die Angaben zu 70% inhaltlich korrekt. Zu 7% waren sie teilweise richtig und zu 22% falsch oder wurden nicht erkannt. Sie waren zu 40% voll-



▶ **Abb. 1** Darstellung der Selektion der Arztbriefe in 2 Stufen als Supportleistung – jede Farbe steht für einen komplett, mittel bis leicht reduzierten Arbeitsaufwand für die Dokumentationskraft. Komplett reduziert ist der Aufwand in den grünen Feldern.

ständig und zu 13% teilweise korrekt eingeordnet. Zu 22% war die Einordnung falsch.

Die UICC-Klassifikation spielte in den von uns ausgewerteten Arztbriefen keine Rolle, sie wird deswegen hier nicht weiter beschrieben.

Die TNM-Klassifikation wurde zu 76% korrekt vorgenommen. In 20% war sie teilweise richtig und zu 2% falsch oder wurde nicht erkannt. Eine korrekte Einordnung fand in 27% der Fälle statt, zu weiteren 22% war sie teils korrekt. Zu 37% war sie nicht korrekt.

▶ **Tab. 4** zeigt die prozentualen Verteilungen der Analyse im Detail.

Diskussion

Die Ableitung der Zeiteinsparungen aus der Stichprobe der Studie ergibt eine Verbesserung um 44%

Um ein ökonomisches Anhaltsmodell für die Bewertung der Kosteneffektivität von LLM-/KI-Systemen in der Tumordokumentation zu entwickeln, wurden verschiedene Aspekte berücksichtigt. Die nachfolgend aufgestellte Kalkulation ist in diesem Sinne als Abschätzung zu verstehen, da mit Hypothesen gerechnet wurde.

Initial wurden die Ergebnisse in ein Schaubild überführt, das es erlaubt, die verschiedenen Aufgaben bei der Informationsextraktion im Sinne einer Reduktion des Arbeitsaufwands zu klassifizieren. Die Überführung der ▶ **Tab. 4** in diese Darstellung zeigt ▶ **Abb. 1**.

▶ **Tab. 4** Prozentuale Verteilung der Richtigkeit der Angaben in den „mit Befund" identifizierten Arztbriefen nach den in ▶ **Tab. 3** definierten 4 Dokumentationsarten. Vollständig verwertbar sind Briefe in der Kategorie 1 B, alle anderen B-Kategorien müssen vom Menschen nachgearbeitet werden.

	Allgemeinbefund		Gleason		UICC		TNM	
Kategorie	A	B	A	B	A	B	A	B
1	43%	43%	70%	40%	0%	0%	76%	27%
2	21%	9%	7%	13%	0%	0%	20%	22%
3	34%	23%	22%	22%	75%	75%	2%	37%
4	2%	25%	1%	25%	25%	25%	2%	14%

Die dunkelgrün bezeichneten Befunde (= „korrekt aussortiert“) können als ein kompletter Wegfall der Screeningarbeit der Briefe klassifiziert werden. Weiterhin kann der hellgrüne Block der Stufe 2 (= „inhaltlich & zeitlich korrekt“) als erhebliche Arbeitserleichterung durch de facto richtig „vorausgefüllte“ Angaben betrachtet werden. Im grau gekennzeichneten Block der Stufe 2 (= „mindestens 1 Faktor falsch“) kann man vereinfachend davon ausgehen, dass keine Erleichterung der Arbeit vorliegt, auch wenn zumindest die relevanten Stellen in den Arztbriefen in Teilen hervorgehoben werden.

Zur Abschätzung kann man pro Patient mit Vorbefunden 10 Seiten Material in Form von 5 Arztbriefen mit je 2 Seiten annehmen. Nimmt man pro Seite 30 Sekunden „Screeningzeit“ an, so würden je 189 Sekunden für die Allgemeine Klassifikation, 162 Sekunden für die Gleason-Klassifikation und 201 Sekunden für die TNM-Klassifikation wegfallen.

Im Bereich der Überführung von Daten aus den identifizierten Arztbriefen kann man von gut 3 Minuten Übertragungszeit pro Befundklasse ausgehen. Vorbefüllte Felder mit markierter „Referenzstelle“ im Quelldokument können hier sehr konservativ gerechnet die Hälfte der Zeit bei der Dokumentation sparen. Übertragen bedeutet dies weitere gesparte 24 Sekunden für die allgemeine Klassifikation, 37 Sekunden für die Gleason-Klassifikation und 20 Sekunden für die TNM-Klassifikation. In Summe würde damit die Arbeit für diesen Teil der Tumordokumentation von statistischen 24 Minuten um 633 Sekunden, also gut 10,5 Minuten reduziert werden. Dies entspräche einer Ersparnis von 44% Zeit pro Dokumentationsvorgang.

Weiterhin ergibt sich aus dem Block der „nicht verwertbaren“ Vordokumentation (grau in ▶ **Abb. 1**) noch Potenzial für Verbesserung im Sinne noch größerer Zeitersparnis. Möglichkeiten ergeben sich dafür vor allem aus der Verbesserung der Prompts in einem Prozess, der als Prompt Engineering bezeichnet wird und im medizinischen Kontext beispielsweise von Meskó beschrieben wird [11]. Auch das sogenannte Fine-Tuning kann als ein Teil davon verstanden werden [12]. Dabei wird dem Programm ein möglichst großer Satz an Daten gegeben, in diesem Fall z. B. 1000 Arztbriefe, in denen die Tumorklassifikation bereits manuell durchgeführt wurde. Die KI erhält so ein besseres „Verständnis“ davon, wie die Informationen am Ende sortiert sein sollen.

Der Aufwand des Dokumentationsvorgangs pro Patient liegt bei knapp 140 € und variiert stark zwischen Kliniken

Eine Studie der Gesellschaft PROGNOS aus dem Jahr 2016 für die deutsche Krebshilfe berechnet für ein typisches onkologisches Spitzenzentrum (CCC) mit 7840 stationären Patienten alleine für den Dokumentationsaufwand pro Jahr eine Summe von 1.094.704 Euro [13], was eine Summe von 139,63 Euro pro Patientenfall für die reinen Dokumentationsaufwände bedeutet.

Vergleicht man diese Erhebung mit anderen Studien z. B. des Bundesministerium für Gesundheit (BMG) aus dem Jahr 2012 im Bereich Mammakarzinom Dokumentation, so zeigt sich eine ähnliche Größenordnung der Kosten für die Dokumentation, wenngleich auch eine enorme Schwankungsbreite aufgezeigt wird. So findet sich in der Studie des BMG eine Tabelle mit der Kostenvarianz von 8 Zentren [14], die im Durchschnitt bei 182 Euro liegen. Für die hier vorliegende Rechnung griff man daher auf die neueste vorliegende Zahl und Studie der PROGNOS zurück und rechnete mit 140 Euro. Auch hier muss berücksichtigt werden, dass Kosten und Prozesse pro Haus deutlich variieren und auch dieser Wert als Abschätzung zu betrachten ist, die sich pro Haus und Prozess unterschiedlich darstellt. Kombiniert man eingesparte Zeit direkt linear mit den hauptsächlich durch Arbeitszeiten des Krankenhauspersonals verursachten Kosten, so kann der wirtschaftliche Impact bei einer Wertschöpfung von 61,54 Euro pro Patient gesehen werden. Da im Bereich der Tumordokumentation ohnehin oftmals Überlastung und/oder Personalmangel vorherrschen, könnte die Wertschöpfung hier gut ohne echten Personalabbau passieren und würde die Effizienz und Qualität der Arbeit ggf. steigern.

Durch LLM-Modelle verursachte Kosten wurden als Pay-per-Use Ansatz gerechnet

Vereinfachend wurde für die Kosten des LLM-Modells ein Betrieb in einer „Pay-per-Use“-Lösung in der Cloud angenommen, wie er aktuell auch stattfindet. Hier kann man im speziellen Fall der LLM-Modelle dann Kosten grob in Input-, Output- und Finetuning-Kosten vornehmen. Die Kosten für Input- und Output sind die Kosten für die Menge der in das System und aus dem System herausgezogenen Datenmengen. Die Kosten für das Finetuning sind komplexer zu berechnen und fallen an, wenn das System mit eigenen Datensätzen fein eingestellt werden soll.

Die Berechnung erfolgte hier anhand der Menge der verarbeiteten Daten in Megabyte (MB) und einer Art eigener Währung, den Tokens. Die Verarbeitung von 1MB an Daten entspricht beispielsweise 34.000 Tokens. Tokens entsprechen wiederum einem Gegenwert in Dollar, der abhängig von der Form der Verarbeitung (Input, Output, Finetuning) und der Programmversion ist.

Da die Outputdatenmenge vernachlässigbar klein ist (im Kilobytebereich) und ein Finetuning nicht stattgefunden hat, werden diese Kosten hier nicht berücksichtigt. Es bleiben damit nur die Kosten für den Input, die wir hier anhand der aktuellen Version von ChatGPT (gpt-4–1106-preview) berechnen.

Darin entsprechen 1000 Tokens einem Wert von 0,01$. Die Verarbeitung eines MB (34K Tokens) kostet also 0,34$. Bei unserem gesamten Datensatz (154 Arztbriefe) in Größe von 26 MB fielen also Kosten von etwa 9$ an. Pro Fall belaufen sich die Kosten für die Zusammenfassung daher auf ca. 0,36$, also ca. 35 Cent. Der Wert kann als vernachlässigbar gegenüber dem geschöpften Wert angesehen werden.

Datenschutz- und Haftungsfragen bergen Herausforderungen

In Bezug auf den medikolegal einwandfreien Einsatz von KI in der Medizin und die damit verbundene Haftung bei Fehlentscheidungen gibt es einige wichtige Literaturreferenzen und Forschungsergebnisse. So wird z. B. diskutiert, wie autonome KI-Systeme in der Medizin verbesserte Ergebnisse versprechen, aber auch Bedenken hinsichtlich Haftung, Regulierung und

Kosten aufwerfen. Ärzte könnten in Fällen von Kunstfehlern bevorzugt werden, wenn sie sich streng an die Empfehlungen validierter KI-Systeme gehalten haben. Allerdings könnten Entwickler von KI-Systemen haftbar gemacht werden, wenn sie bei der Entwicklung und Implementierung nicht den branchenüblichen Best Practices folgen [15].

Die Weiterentwicklung von KI-Modellen führt dabei aber auch zu zunehmend komplexen Haftungsfragen. Autonome KI-Systeme können komplexe medizinische Aufgaben übernehmen und verwischen dabei die Grenze zwischen menschlicher und KI-Entscheidungsfindung. Fehler, die zu Patientenschäden führen, sind unvermeidlich. Ergebnisse oder Empfehlungen von vollständig autonomen KI-Systemen können von medizinischem Personal verwendet oder überprüft werden, was potenziell dann aber eine Haftung für Kliniker mit sich bringt [16].

Von Rechtswissenschaftlern wurden in einer Feldstudie dazu Szenarien identifiziert, in denen ein Arzt ein autonomes KI-System nutzt und die relevant für die Haftungsfrage sind [17]. Es gibt demnach nur 2 Szenarien, in denen der Arzt möglicherweise haftbar gemacht werden könnte: (A) wenn das System eine Behandlung empfiehlt, die dem aktuellen Standard entspricht, der Arzt diese Empfehlung jedoch ignoriert und dadurch dem Patienten Schaden zufügt; und (B) wenn das System eine Behandlung empfiehlt, die nicht dem Standard entspricht, der Arzt dieser Empfehlung folgt und dadurch dem Patienten Schaden zufügt. Es wird erwartet, dass in diesen Szenarien Richter und Geschworene die Ärzte bevorzugen, wenn sie den Empfehlungen der KI folgen. Zuletzt könnten auch die Schöpfer der KI-Systeme für fahrlässiges Design oder Implementierung des KI-Systems haftbar gemacht werden, wenn dies zu Patientenschäden führt. Wenn beispielsweise der KI-Entwickler die KI nicht gemäß den branchenüblichen Best Practices validiert, könnte er für die durch Fahrlässigkeit verursachten Patientenschäden haftbar gemacht werden. All diese Aspekte werden auch unter verschiedenen Aspekten der Ethik ausführlich in der Literatur diskutiert und empfehlen meist eine Abwägung zwischen den Risiken eines Einsatzes vs. dem Nichteinsatz der Technologie [18].

Die Recherche im Rahmen der Studie verdeutlicht, dass der Einsatz von KI in der Medizin nicht nur technische, sondern auch erhebliche rechtliche und ethische Herausforderungen mit sich bringt, insbesondere im Hinblick auf die Haftung bei Fehlentscheidungen. Diese Aspekte müssen sorgfältig berücksichtigt werden, um einen medikolegal einwandfreien Einsatz von KI in der Medizin zu gewährleisten.

Bewertung des Performance-Tests – Potenziale von LLMs für die Vereinfachung von medizinischen Dokumentationsvorgängen

Zusammenfassend gesprochen stellt die Integration von ChatGPT und ähnlichen LLMs im medizinischen Bereich, insbesondere bei der Tumordokumentation, einen vielversprechenden Weg dar. Diese Integration könnte Effizienz, Genauigkeit und die Gesamtqualität der medizinischen Dokumentation verbessern.

Der aus der Stichprobe abgeleitete Hebel von 35 Cent Kosten für eine 60-Euro-Wertschöpfung scheint enorm und muss daher sicherlich kritisch betrachtet werden.

In der Realität besteht der Dokumentationsvorgang aus mehreren Schritten und nicht nur der reinen Übertragung aus gescreenten Vorbefunden. Daher ist die Zeitersparnis nur teils auf den gesamten Vorgang zu projizieren. Auf der anderen Seite wurde mit einem nahezu untrainierten System gearbeitet, das noch nicht auf die spezifischen medizinischen Daten ausgerichtet wurde und auch noch nicht an den Arbeitsablauf der Dokumentation adaptiert wurde.

Wichtig scheint es auch, zu verstehen, dass die Bearbeitung mit einem LLM anders abläuft als eine reine „Textsuchfunktion“. Das LLM ist wesentlich höherwertig in der Lage, aus Informationen wie „Tumor hat gestreut“, die richtige TNM-Klassifikation abzuleiten. Hier veranschaulicht die Arbeit an den Prompts in dieser Arbeit auch, dass man mit Prompts teils nur einen groben Kontext vorgeben muss. So zeigt das Beispiel der Klassifikation in Gleason und TNM, die der KI anfangs nur oberflächlich beschrieben wurden und trotzdem dann auch in der Tiefe weitgehend richtig bearbeitet wurden, dass das Modell auch aus seinen vortrainierten Informationen Bezüge richtig herstellen und ergänzen kann. Auf der anderen Seite kann aber auch eine reine Suchfunktion besser sein, wenn es um bestimmte eindeutige Suchwörter geht, da das LLM eben auch „ungenau“ arbeitet und teilweise Wörter im Kontext auch missinterpretiert. Hier scheint es wichtig zu verstehen, dass ein optimiertes System daher auch aus einer Kaskade von Abfragen – ggf. sogar Textsuchfunktionen, Prompts und LLM-Umformungen mehrfach kombiniert – bestehen kann. LLM-Systeme der Zukunft könnten so aus einer Stufung von Informationsprozessierungen und sogar mehreren „Agenten“, unter denen am Ende das beste oder richtige Ergebnis vereinbart wird, bestehen.

Grundsätzlich besteht also sicher mit gegebenem weiterem Aufwand großes Potenzial für das Feld der Tumordokumentation. Weiterhin ist in der Arbeit der gesamte Aspekt des Finetunings des Modells sowie die Einbettung einer echten „Software“ in den Arbeitsablauf gar nicht betrachtet worden. Vereinfachend könnte man sagen, dass hierfür 61,19 Euro pro Dokumentationsvorgang abzüglich einer erwarteten Marge zur Verfügung stehen würden. Die Autoren gehen weiterhin davon aus, dass die ggf. zu gut berechneten Effizienzgewinne durch die Projektion des Dokumentationsnutzens auf den gesamten Vorgang mindestens durch die hier skizzierten noch denkbaren Verbesserungen eines spezifisch angepassten Systems möglich sind. Die gefundene und projizierte Größenordnung deckt sich außerdem mit Größenordnungen aus Arbeiten in anderen Feldern [19], sodass man hier von einer richtigen Annahme ausgeht.

Der Einsatz von generativer KI im Kontext von Such-, Übertragungs- und Interpretationsarbeiten bei der Erstellung einer Tumordokumentation ist daher als ein vielversprechender Ansatz zu sehen. Die hier vorgelegte Arbeit kann aber nur als erster Schritt einer Bewertung gesehen werden – es gilt nun mit in den Ablauf eingebetteten Ansätzen im echten klinischen Betrieb herauszufinden, ob die Interaktion zwischen Dokumen-

tar*innen und LLM-Modellen in der Realität Prozesse beschleunigt. Um dies vorzunehmen, erfordert es eine sorgfältige Navigation durch Datenschutz- und Verantwortungsfragen. Die zukünftige Forschung und Anwendung solcher Systeme sollten sich weiterhin auf die Entwicklung sicherer, ethischer und effektiver Ansätze zur Integration von KI in die medizinische Praxis konzentrieren.

Die hier vorliegende Arbeit könnte dabei ggf. noch einen weiteren wichtigen Aspekt speisen. So ist bei der Weiterentwicklung bzw. Erprobung anderer Modelle für medizinische Use-Cases auch immer eine Schlüsselfrage, welche Benchmarks am Ende genutzt werden, um die Performance eines Systems zu messen. Bisher dient dazu meist das medizinische US-Staatsexamen. Denkbar wäre jedoch auch für spezifischere Aufgaben wie die Tumordokumentation die Verwendung von Testdatensätzen zu diesem Zweck. Die hier generierten Daten könnten auch dazu verwendet werden und eine Art „Benchmark-Datenset" bilden.

Die aktuellen Arbeiten z. B. der IT-Abteilung des Universitätsklinikums Schleswig-Holstein, LLM-Modelle in einer sicheren Cloud-Umgebung durch getrennte Schlüsselverwaltung bei der Telekom anzugehen, könnten den Part des Datenschutzes für die Verarbeitung von Echtdaten gut lösen [20]. Für die Einbettung eines LLMs in den medizinischen Dokumentationsablauf müssten dann noch einfache Werkzeuge und Arbeitsoberflächen sowie verfeinerte gestufte Abfragen (Prompts) [21] an die Systeme in Betrieb genommen werden, um einen Pilotbetrieb in klinischer Umgebung zu ermöglichen. Hier muss sich dann zeigen, wie das hier erhobene theoretische Potenzial in einem echten Betrieb zu heben ist und ob diese ersten Abschätzungen aus einer Stichprobe auch die Effizienzgewinne in der praktischen klinischen Realität spiegeln.

Interessenkonflikt

Die Autorinnen/Autoren geben an, dass kein Interessenkonflikt besteht.

Literatur

[1] LLMs: How They Work and Their Applications | ML6. Zugriff am 22. Februar 2024: https://www.ml6.eu/resources/large-language-models

[2] Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: An emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks Stand Eval 2023; 3: 100105 doi: 10.1016/j.tbench.2023.100105

[3] Hadi MU, Al-Tashi Q, Qureshi R et al. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. 2023

[4] Kung TH, Cheatham M, Medenilla A et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2: e0000198 doi: 10.1371/journal.pdig.0000198

[5] Ali SR, Dobbs TD, Hutchings HA et al. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023; 5: e179–e181 doi: 10.1016/S2589-7500(23)00048-1

[6] Johnson D, Goodman R, Patrinely J et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq 2023: doi: 10.21203/rs.3.rs-2566942/v1

[7] Rumbold JMM, Pierscionek B. The Effect of the General Data Protection Regulation on Medical Research. J Med Internet Res 2017; 19: e47 doi: 10.2196/jmir.7108

[8] Walonoski J, Hall D, Bates K et al. The "Coherent Data Set": Combining Patient Data and Imaging in a Comprehensive, Synthetic Health Record. Electronics 2022; 11: 1199 doi: 10.3390/electronics11081199

[9] Hernandez M, Epelde G, Alberdi A et al. Synthetic data generation for tabular health records: A systematic review. Neurocomputing 2022; 493: 28–45 doi: 10.1016/j.neucom.2022.04.053

[10] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019; 25: 44–56 doi: 10.1038/s41591-018-0300-7

[11] Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res 2023; 25: e50638 doi: 10.2196/50638

[12] The Ultimate Guide to LLM Fine Tuning: Best Practices & Tools | Lakera – Protecting AI teams that disrupt the world. Zugriff am 22. Februar 2024: https://www.lakera.ai/blog/llm-fine-tuning-guide

[13] Prognos_Endbericht_Deutsche_Krebshilfe.pdf

[14] 140206_Abschlussbericht_Projekt_Mammakarzinom.pdf

[15] Saenz AD, Harned Z, Banerjee O et al. Autonomous AI systems in the face of liability, regulations and costs. Npj Digit Med 2023; 6: 1–3 doi: 10.1038/s41746-023-00929-1

[16] Price WN, Gerke S, Cohen IG. Potential Liability for Physicians Using Artificial Intelligence. JAMA 2019; 322: 1765–1766 doi: 10.1001/jama.2019.15064

[17] Tobia K, Nielsen A, Stremitzer A. When Does Physician Use of AI Increase Liability? J Nucl Med Off Publ Soc Nucl Med 2021; 62: 17–21 doi: 10.2967/jnumed.120.256032

[18] Tang L, Li J, Fantus S. Medical artificial intelligence ethics: A systematic review of empirical studies. Digit Health 2023; 9: 20552076231186064 doi: 10.1177/20552076231186064

[19] Zarifhonarvar A. Economics of ChatGPT: A Labor Market View on the Occupational Impact of Artificial Intelligence. 2023

[20] Ehrgeiziger Plan: UKSH geht in die Cloud von Telekom und Google. kma Online. 2024: Zugriff am 22. Februar 2024: https://www.kma-online.de/aktuelles/it-digital-health/detail/uksh-geht-in-die-cloud-von-telekom-und-google-51383

[21] Hughes A. The power of prompting. Microsoft Res 2023. Zugriff am 22. Februar 2024: https://www.microsoft.com/en-us/research/blog/the-power-of-prompting/