Experten-angeleitet und domänenspezifisch: MEREDITH

LLM-Systeme als klinische Entscheidungshilfen in der Präzisionsonkologie

19. November 2024
MIRACUM DIFUTURE Kolloquium

Medizinische Abkürzung
MTB: Molekulares Tumorboard

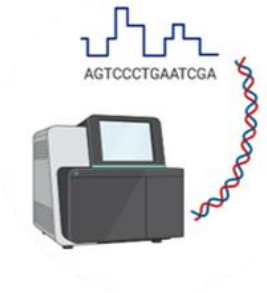
Dr. med. Jacqueline Lammert
Clinician Scientist, Frauenklinik

Molekulares Tumorboard (MTB) am ZPM^{TUM}

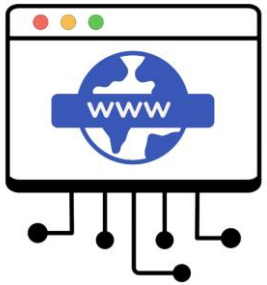
Ablauf

Pat.einwilligung und -einschluss

Gewinnung/ Anforderung von Tumormaterial



Molekulare Diagnostik, Bioinformatik



Klinische Interpretation der Molekularen Daten

Falldiskussion im Molekularen Tumorboard (MTB)

Zielgerichtete Therapie
- Klinische Studie
- Off-Label

Automatisierung mit *generativer KI*

Präzisionsonkologische Literatursynthese ist aufwändig

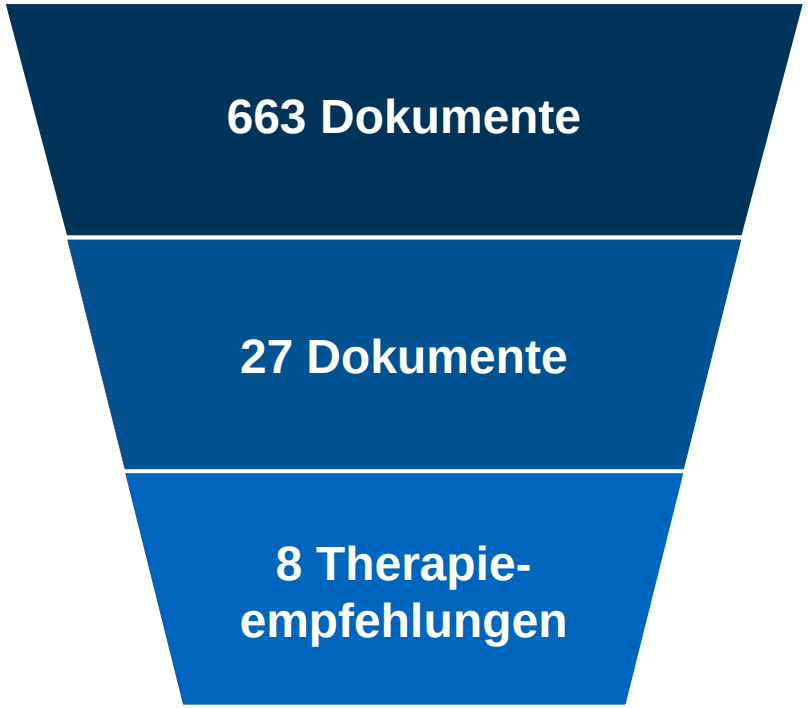
Ein reales Beispiel

73-jährige Patientin mit uterinem Karzinosarkom, metastasiert, Zustand nach 2L Chemotherapie

Sichtung der Quellen auf PubMed, Leitlinien, ClinicalTrials.gov, EMA, FDA,…

Detaillierte Analyse passender Quellen

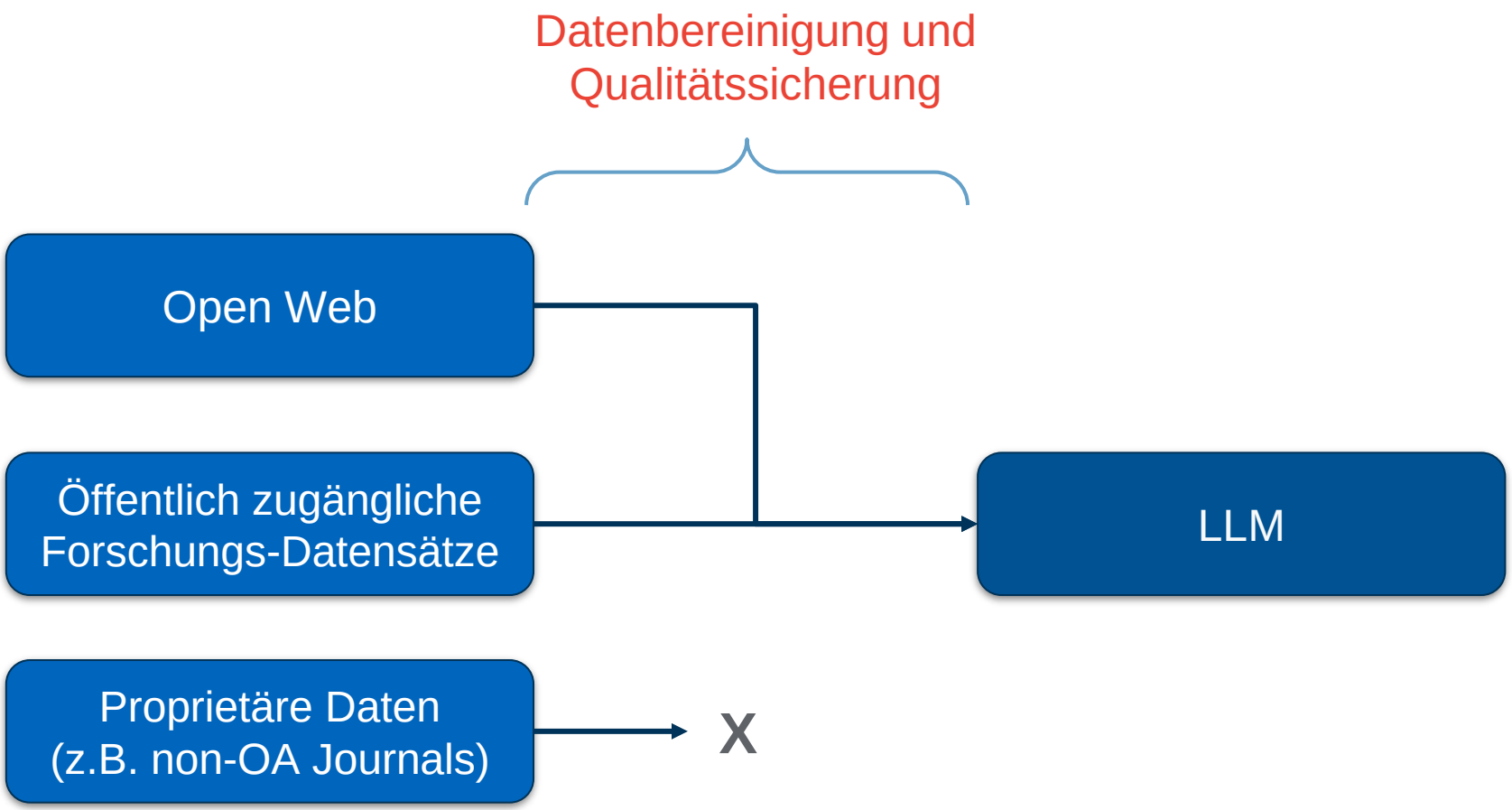
Synthese individualisierter Therapieempfehlungen



Quelle: Lammert *et al.* (2024), https://arxiv.org/abs/2409.00544

Große Sprachmodelle (LLM) können Wissenschaftler*innen darin unterstützen, medizinische Fachliteratur automatisiert zu verarbeiten.

Öffentliche LLMs wie ChatGPT werden an öffentlich verfügbarem Wissen trainiert.



Das bedeutet, dass einfache LLMs (z.B. ChatGPT) nicht gut auf komplexe medizinische Fragen reagieren können, weil sie nicht in diesem Kontext ausgebildet sind (fehlende Domänenspezifika).

Original Investigation | Oncology

November 17, 2023

Leveraging Large Language Models for Decision Support in Personalized Oncology

Manuela Benary, PhD[1,2]; Xing David Wang, MSc[3]; Max Schmidt, MD[1,4]; et al

» Author Affiliations | Article Information

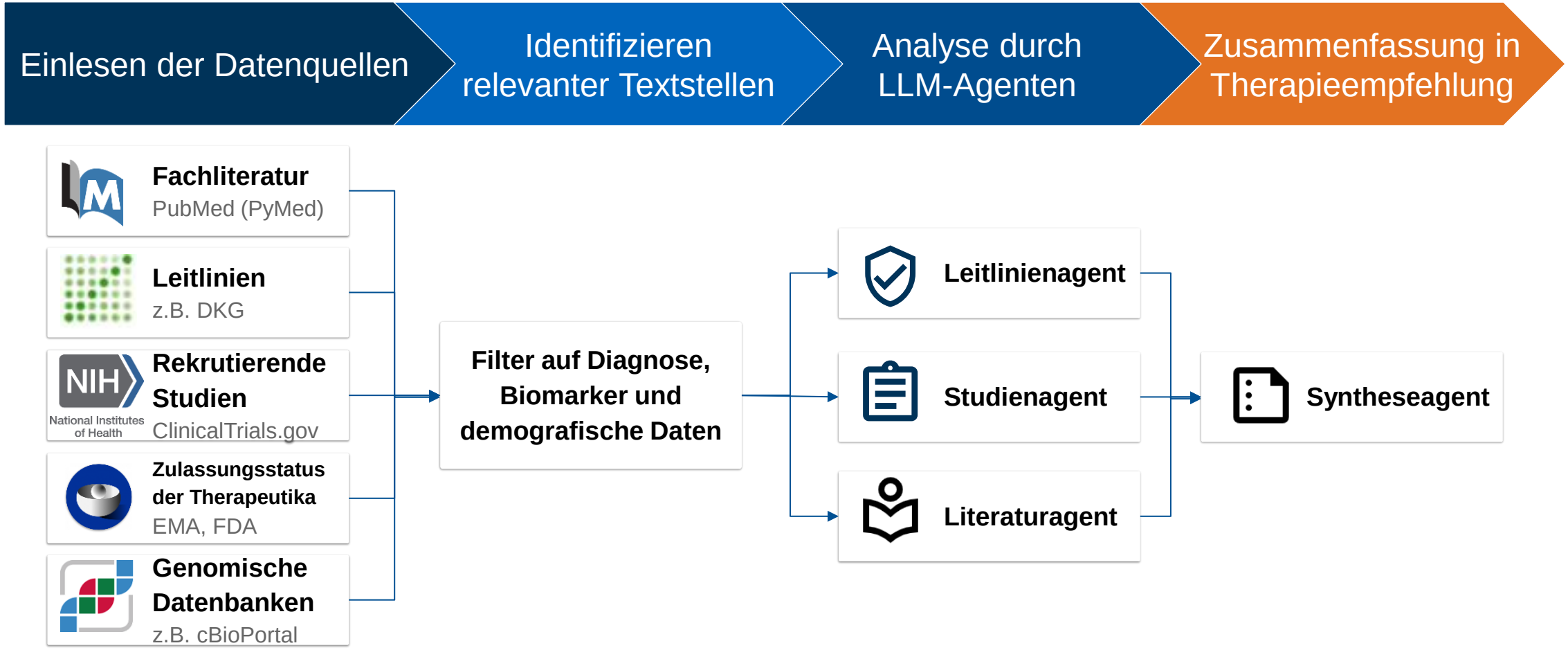
JAMA Netw Open. 2023;6(11):e2343689. doi:10.1001/jamanetworkopen.2023.43689

Conclusions and Relevance In this diagnostic study, treatment options of LLMs in precision oncology did not reach the quality and credibility of human experts; however, they generated helpful ideas that might have complemented established procedures. Considering technological progress, LLMs could play an increasingly important role in assisting with screening and selecting relevant biomedical literature to support evidence-based, personalized treatment decisions.

Quelle: Benary *et al.* (2024), https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2812097

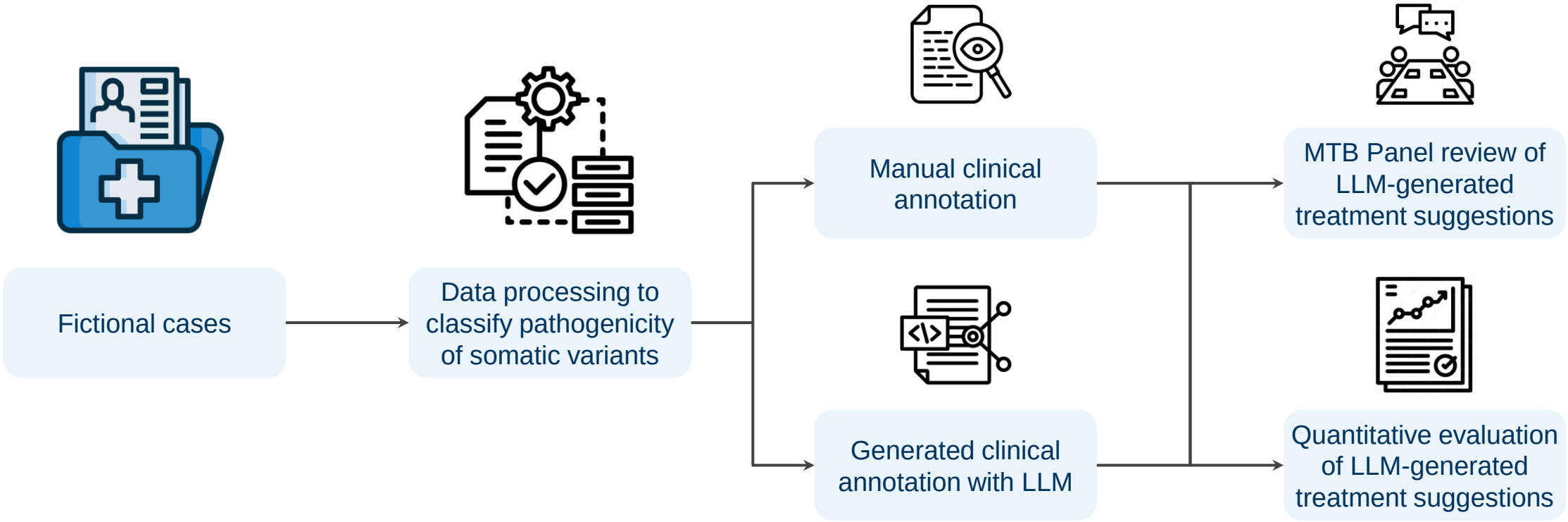
Haben domänenspezifische LLMs das Potential, uns in der klinischen Interpretation von molekularen Daten zu unterstützen?

MEREDITH repliziert die Vorgehensweise der MTB^{TUM}-Experten in der klinischen Interpretation von molekularen Daten.



Quelle: Lammert *et al.* (2024), https://doi.org/10.1200/PO-24-00478

Um die LLM-generierten Therapieoptionen einzuordnen, führten wir eine qualitative Analyse durch MTB^{TUM}-Experten und eine quantitative Analyse mittels Cosinus-Ähnlichkeit der Textvektor-Einbettungen durch.



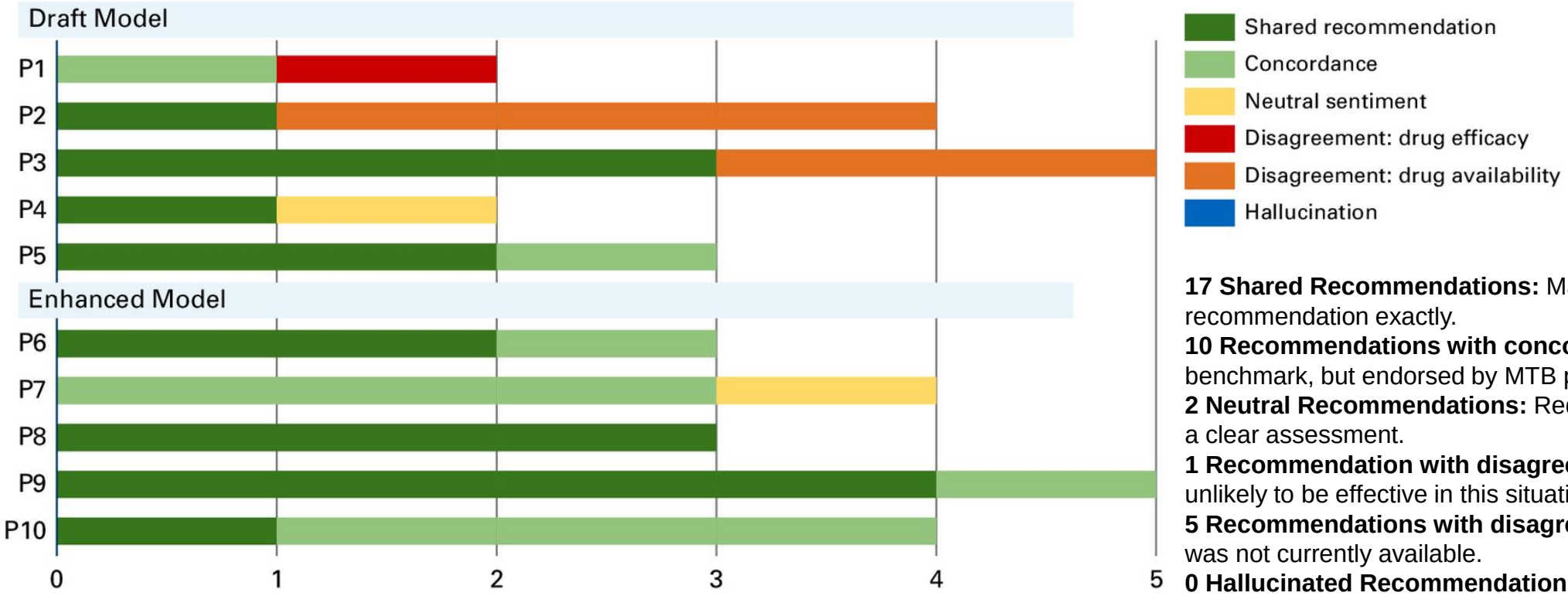
Quelle: Lammert *et al.* (2024), https://doi.org/10.1200/PO-24-00478

Das LLM-System wurde so konfiguriert, dass es ein breiteres Spektrum an Behandlungsoptionen vorschlug. Die anschließende Überprüfung und Priorisierung durch menschliche Expert*innen stellte die Patientenrelevanz sicher.

Patient	Medical experts	Draft system[a]	Enhanced system[b]
1	1	2	2
2	1	4	3
3	4	5	6
4	2	2	4
5	2	3	5
6	2	2	3
7	1	4	4
8	2	2	3
9	3	4	5
10	3	4	4

[a] The draft LLM system was discussed by medical experts for cases 1-5 only.
[b] The enhanced LLM system was discussed by medical experts for cases 6-10 only.

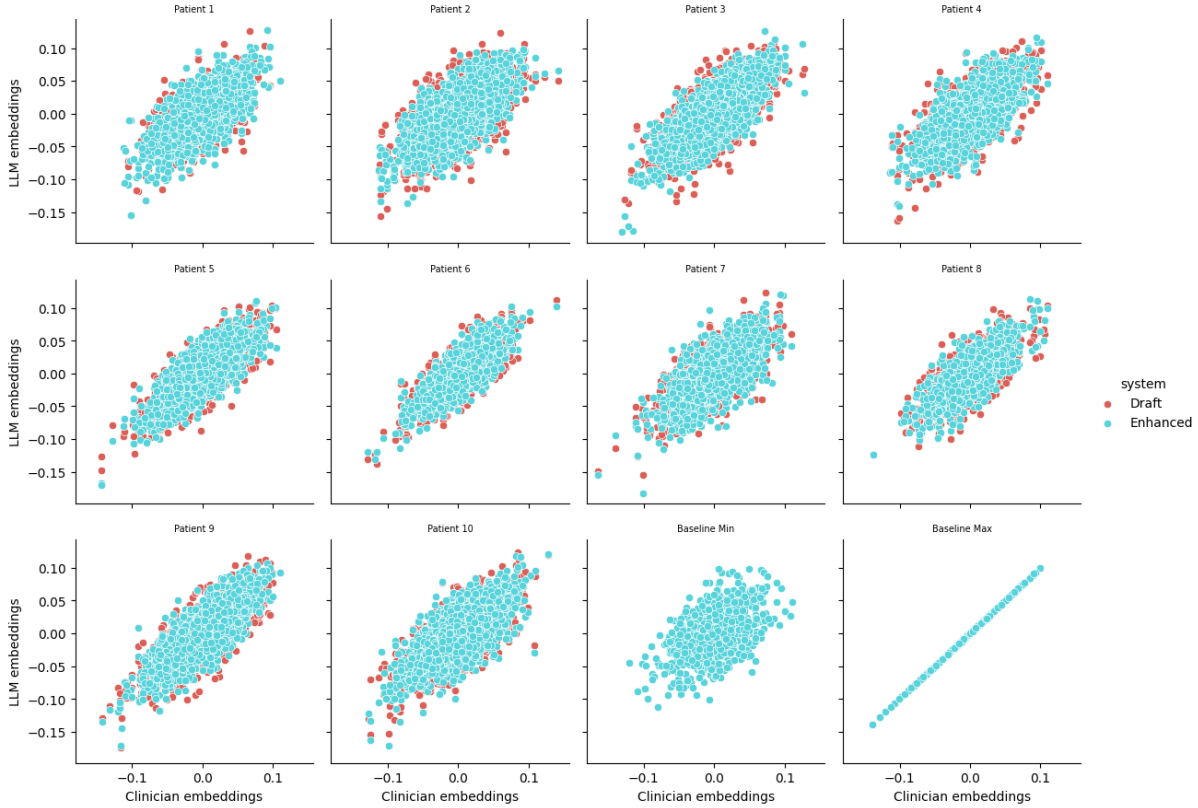
Die Analyse ergab eine hohe Übereinstimmung (94,7%) zwischen LLM-generierten Therapieoptionen und den Empfehlungen menschlicher MTB-Experten (qualitative Einschätzung).



17 Shared Recommendations: Matched expert recommendation exactly.
10 Recommendations with concordance: Unmentioned in benchmark, but endorsed by MTB panel.
2 Neutral Recommendations: Required further information for a clear assessment.
1 Recommendation with disagreement (Efficacy): Drug unlikely to be effective in this situation.
5 Recommendations with disagreement (Availability): Drug was not currently available.
0 Hallucinated Recommendations: None of the LLM options were hallucinated, all references were cited correctly.

Quelle: Lammert *et al.* (2024), https://doi.org/10.1200/PO-24-00478

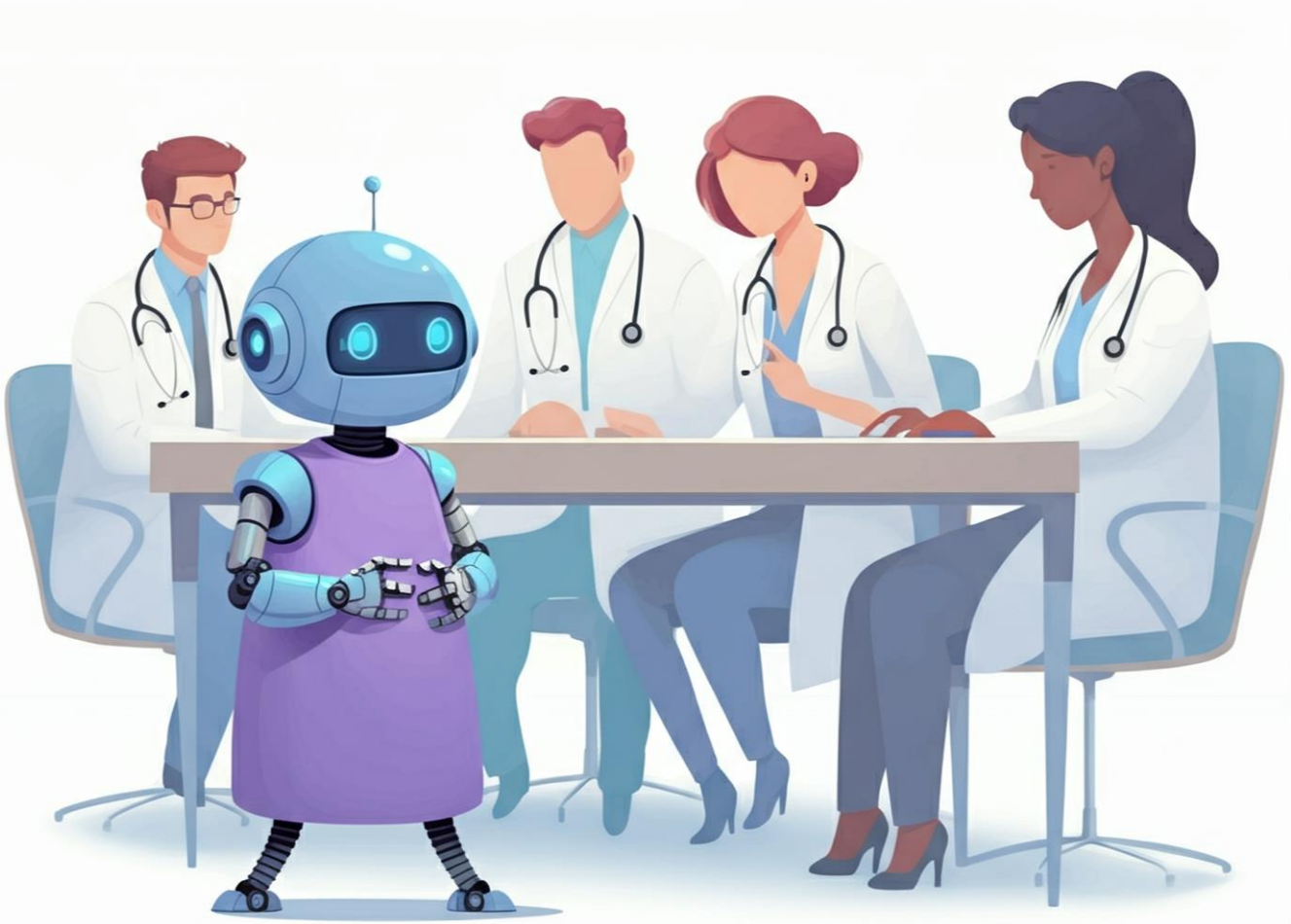
Die semantische Ähnlichkeit zwischen LLM und menschlichen Experten lag im Bereich der für MTBs beschriebenen Interraterreliabilität (quantitative Einschätzung).



- For patients x=[1, 10], the mean cosine similarity m(cos(θ) between MTB expert recommendation and iteration 1 of the LLM system was calculated as m(cos(θ)=0.71, IQR [0.61, 0.80].
- For patients x=[1,10], the mean cosine similarity m(cos(θ) between MTB expert recommendation and iteration 2 of the LLM system was calculated as m(cos(θ)=0.76, IQR [0.73, 0.82].
- That means cosine similarity increased by an average of 9% from iteration 1 to iteration 2.
- The paired *t*-test showed this increase to be significant with p = 0.01.

Quelle: Lammert *et al.* (2024), https://doi.org/10.1200/PO-24-00478

Wie können wir LLM-Systeme in den realen Versorgungsalltag integrieren?



Nur 5% der Publikationen zum Einsatz von LLMs in der Medizin basieren auf realen Versorgungsdaten.

JAMA. 2024 Oct 15:e2421700. doi: 10.1001/jama.2024.21700. Online ahead of print.

Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review

Suhana Bedi [1], Yutong Liu [2], Lucy Orr-Ewing [2], Dev Dash [2 3], Sanmi Koyejo [4], Alison Callahan [3], Jason A Fries [3], Michael Wornow [3], Akshay Swaminathan [3], Lisa Soleymani Lehmann [5], Hyo Jung Hong [6], Mehr Kashyap [7], Akash R Chaurasia [3], Nirav R Shah [2], Karandeep Singh [8], Troy Tazbaz [9], Arnold Milstein [2], Michael A Pfeffer [10], Nigam H Shah [2 3]

Affiliations + expand

PMID: 39405325 PMCID: PMC11480901 (available on 2025-04-15) DOI: 10.1001/jama.2024.21700

Results: Of 519 studies reviewed, published between January 1, 2022, and February 19, 2024, only 5% used real patient care data for LLM evaluation. The most common health care tasks were assessing medical knowledge such as answering medical licensing examination questions (44.5%) and making diagnoses (19.5%). Administrative tasks such as assigning billing codes (0.2%) and writing prescriptions (0.2%) were less studied. For NLP and NLU tasks, most studies focused on question answering (84.2%), while tasks such as summarization (8.9%) and conversational dialogue (3.3%) were infrequent. Almost all studies (95.4%) used accuracy as the primary dimension of evaluation; fairness, bias, and toxicity (15.8%), deployment considerations (4.6%), and calibration and uncertainty (1.2%) were infrequently measured. Finally, in terms of medical specialty area, most studies were in generic health care applications (25.6%), internal medicine (16.4%), surgery (11.4%), and ophthalmology (6.9%), with nuclear medicine (0.6%), physical medicine (0.4%), and medical genetics (0.2%) being the least represented.

Kliniker*innen und Forschende führen im Wesentlichen drei Hürden bei der Implementierung an.

Technologie

Unzureichende IT-Infrastruktur: Medizinische Einrichtungen sind angehalten, ihre eigene Infrastruktur aufzubauen, ohne genügend geschultes Personal für deren Betrieb zu haben.

Interoperabilität

Daten werden zwischen Gesundheitseinrichtungen oft in unstrukturierter Form oder sogar auf Papier ausgetauscht; es gibt zwar Rahmen und Standards, jedoch fehlt die Breitenanwendung, z.B. MII Broad Consent v1.7.2.

Nutzer-Akzeptanz

Mehr als 80% der Deutschen glauben, dass KI die Gesundheitsversorgung verbessern wird, dennoch werden digitale und KI-gestützte Tools im Gesundheitswesen nur langsam angenommen.

Fokus für heute

Quelle: OECD (2024), https://doi.org/10.1787/609808d6-en

Drei Schlüsselfaktoren, um die Herzen und Köpfe der Nutzer*innen zu gewinnen

- Ein nutzerzentriertes Design entwickeln
- Ein relevantes Problem identifizieren und lösen
- Erklärbarkeit und Transparenz sicherstellen

Design Thinking: Den Anwender in den Mittelpunkt stellen

Hallo, mein Name ist

Jacqueline

Ich bin

Klinikärztin

Meine Fähigkeiten und Erfahrungen

- Vertraut mit molekularpathologischen Daten und deren Translation in Therapiepläne.
- Grundverständnis für Technik im Allgemeinen und medizinische Software im Besonderen.
- Offen für die Einführung neuer Technologien, die die Patientenversorgung und die Effizienz verbessern.

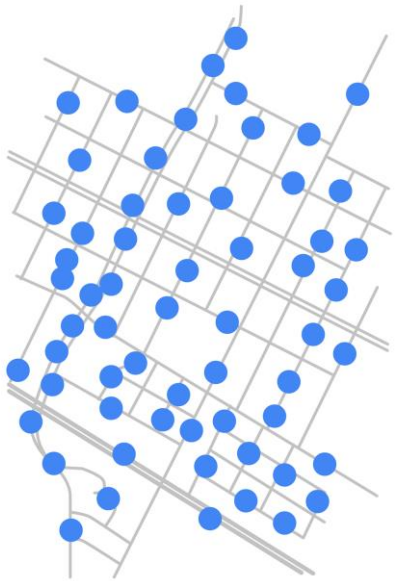
Meine Ziele sind

- die bestmögliche Behandlung für meine Patientinnen auf Grundlage ihres Biomarkerprofils bereitzustellen.
- über die neuesten klinischen Studien in der Präzisionsonkologie auf dem Laufenden zu bleiben.
- eine Verbesserung der Behandlungsergebnisse und der Lebensqualität der Patientinnen.

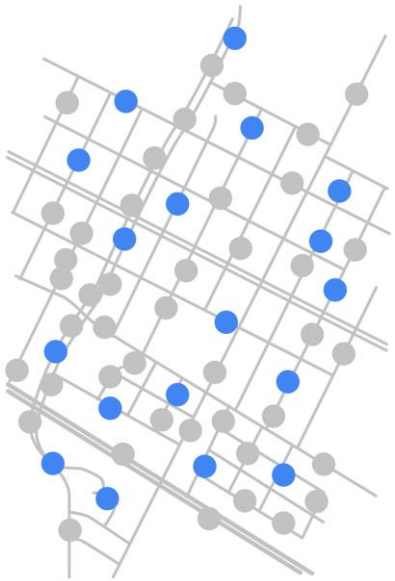
Meine größten Schmerzpunkte sind

- begrenzte Zeit für gründliche Recherche und Evaluation aller potenziellen Behandlungsoptionen.
- fehlende Verfügbarkeit von Daten in strukturierten Formaten, die ich leicht analysieren kann.
- hoher Verwaltungsaufwand und manuelle Dokumentation statt Patientenversorgung.

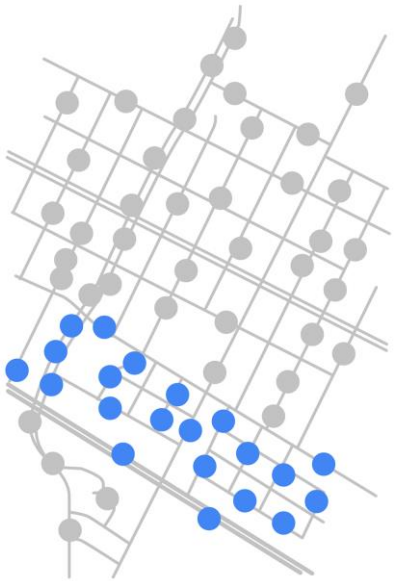
Nicht auf jedes einzelne Schlagloch konzentrieren – Fokus auf den kritischen Pfad



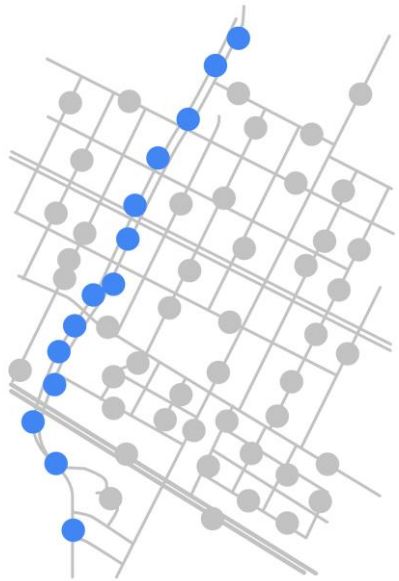
Schlaglöcher auf dem Weg



Häufigste berichtete Probleme



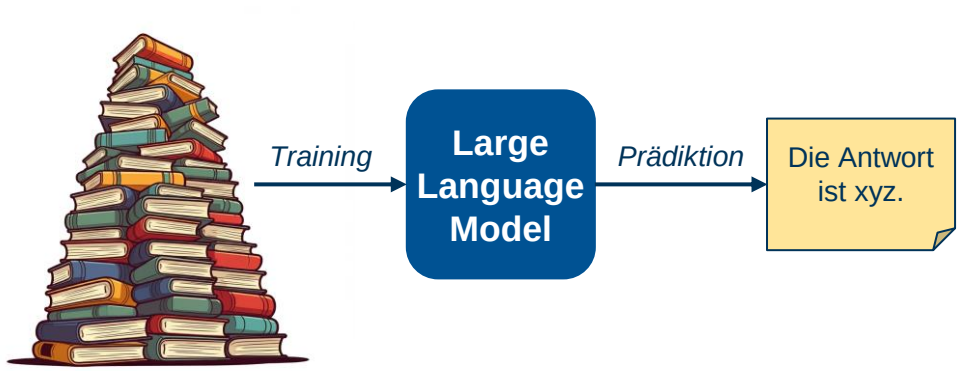
Eine Reihe von Merkmalen



Ein Pfad ✓

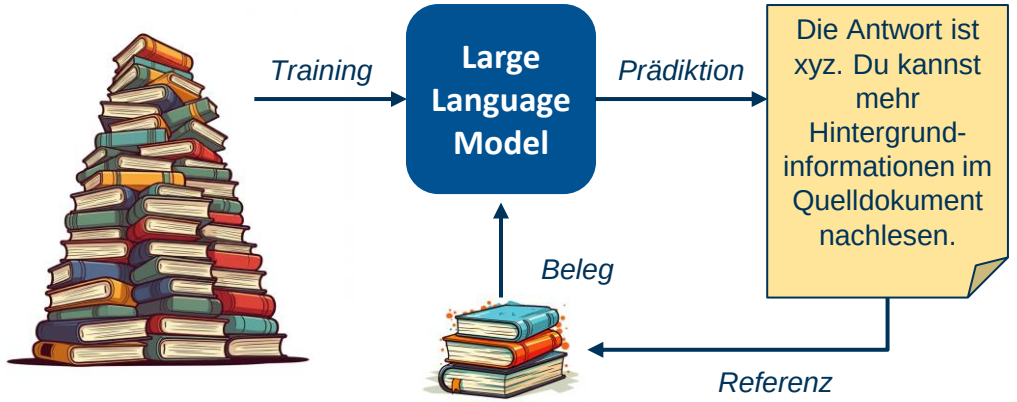
Erstellen von nachvollziehbaren Ergebnissen, um Transparenz zu fördern und Vertrauen zu gewinnen.

In der Regel kann KI ihre Outputs nicht erklären.



- KI-Modelle werden mit einer großen Menge an gelabelten Daten trainiert.
- Sie verwenden mathematische Modelle, um die wahrscheinlichste Antwort („Label“) auf eine Frage („Feature“) vorherzusagen.
- In der Regel sind diese Vorhersagen für den Anwender nicht erklärbar.

Durch Referenzieren von Quelldokumenten können wir die zugrunde liegende Argumentation nachvollziehbar machen.



- Wir integrieren relevante Daten (z. B. eine medizinische Leitlinie) als Grundlage und instruieren das Modell, diese für Vorhersagen zu nutzen.
- Das Modell kann seine Quellen zitieren und referenzieren, wodurch die Nachvollziehbarkeit für Nutzer steigt und die Wahrscheinlichkeit von „Halluzinationen“ sinkt.


“Technology is built by humans and controlled by humans, and we cannot talk about technology as an independent agent acting outside of human decisions and accountability – this is true for AI as much as anything else.”
Meredith Whittaker
llamert

Herzlichen Dank an alle, welche an der Entwicklung und Validierung von MEREDITH mitgewirkt haben!

Bereitstellung der 10 onkologischen Fallvignetten:
Charité Berlin
Manuela Benary
Damian Rieke

Replikation eines echten MTB-Workflows:
ZPM^{TUM}-MTB
Anna Lena Illert
Anna Durner
Johannes Jung
Sebastian Lange
Alisa M. Lörsch
Carolin Mogler
Nicole Pfarr
Ulrich A. Schatz
Kristina Schwamborn
Christof Winter

Ethische Implikationen:
Ethikkommission MRI/TUM
Sonja Mathes

Klinisch-wissenschaftlicher Input:
Radioonkologie/Frauenklinik MRI/TUM
Kai J. Borm
Tobias Dreyer
Marion Kiechle

Technischer Input:
EKF Center for Digital Health, TU Dresden
Dyke Ferber
Jakob Nikolas Kather
AI Engineering, Google Cloud, München
Leonid Kuligin
Max Tschochohei

Vielen Dank!

Jacqueline.Lammert@mri.tum.de