communications medicine

Article

https://doi.org/10.1038/s43856-024-00717-2

Current applications and challenges in large language models for patient care: a systematic review

Felix Busch[1] ✉, Lena Hoffmann[2], Christopher Rueger[2], Elon HC van Dijk[3,4], Rawen Kader[5], Esteban Ortiz-Prado[6], Marcus R. Makowski[1], Luca Saba[7], Martin Hadamitzky[8], Jakob Nikolas Kather[9,10], Daniel Truhn[11], Renato Cuocolo[12], Lisa C. Adams[1,13] & Keno K. Bressem[1,8,13]

Abstract

Background The introduction of large language models (LLMs) into clinical practice promises to improve patient education and empowerment, thereby personalizing medical care and broadening access to medical knowledge. Despite the popularity of LLMs, there is a significant gap in systematized information on their use in patient care. Therefore, this systematic review aims to synthesize current applications and limitations of LLMs in patient care.

Methods We systematically searched 5 databases for qualitative, quantitative, and mixed methods articles on LLMs in patient care published between 2022 and 2023. From 4349 initial records, 89 studies across 29 medical specialties were included. Quality assessment was performed using the Mixed Methods Appraisal Tool 2018. A data-driven convergent synthesis approach was applied for thematic syntheses of LLM applications and limitations using free line-by-line coding in Dedoose.

Results We show that most studies investigate Generative Pre-trained Transformers (GPT)-3.5 (53.2%, $n = 66$ of 124 different LLMs examined) and GPT-4 (26.6%, $n = 33/124$) in answering medical questions, followed by patient information generation, including medical text summarization or translation, and clinical documentation. Our analysis delineates two primary domains of LLM limitations: design and output. Design limitations include 6 second-order and 12 third-order codes, such as lack of medical domain optimization, data transparency, and accessibility issues, while output limitations include 9 second-order and 32 third-order codes, for example, non-reproducibility, non-comprehensiveness, incorrectness, unsafety, and bias.

Conclusions This review systematically maps LLM applications and limitations in patient care, providing a foundational framework and taxonomy for their implementation and evaluation in healthcare settings.

Plain Language Summary

Large language models (LLMs) are computer programs that can generate human-like text. They promise to improve patient education and expand access to medical information by helping patients better understand health conditions and treatment options. However, more information is needed about how these tools are used in patient care and the challenges they present. In this review, researchers analyzed 89 studies from 2022 to 2023 covering 29 medical specialties. These studies explored ways LLMs are used: for example, answering patient questions, summarizing or translating medical texts, and supporting clinical paperwork. While these tools show potential, the review highlights limitations. Many LLMs are not optimized for medical use, lack transparency about data use, and can be difficult for some users to access. Additionally, the text they generate may sometimes be inaccurate, incomplete, or biased, raising safety concerns.

Public and academic interest in large language models (LLMs) and their potential applications has increased substantially, especially since the release of OpenAI's ChatGPT (Chat Generative Pre-trained Transformers) in November 2022[1–3]. One of the main reasons for their popularity is the remarkable ability to mimic human writing, a result of extensive training on massive amounts of text and reinforcement learning from human feedback[4].

Since most LLMs are designed as general-purpose chatbots, recent research has focused on developing specialized models for the medical domain, such as Meditron or BioMistral, by enriching the training data of LLMs with medical knowledge[5,6]. However, this approach to fine-tuning LLMs requires significant computational resources that are not available to everyone and is also not applicable to closed-source LLMs, which are often the most powerful. Therefore, another approach to improve LLMs for

A full list of affiliations appears at the end of the paper. ✉e-mail: felix.busch@tum.de

biomedicine is to use techniques such as Retrieval-Augmented Generation (RAG)[7]. RAG allows information to be dynamically retrieved from medical databases during the model generation process, enriching the output with medical knowledge without the need to train the model.

LLMs hold great promise for improving the efficiency and accuracy of healthcare delivery, e.g., by extracting clinical information from electronic health records, summarizing, structuring, or explaining medical texts, streamlining administrative tasks in clinical practice, and enhancing medical research, quality control, and education[8–10]. In addition, LLMs have been shown to be versatile tools for supporting diagnosis or serving as prognostic models[11,12].

However, despite the growing body of research and the clear potential of LLMs in healthcare, there is a gap in terms of systematized information towards their use in patient care (i.e., the use of LLMs by patients or their caregivers for disease management and support). In contrast to applications primarily aimed at healthcare professionals, LLMs in patient care could be used for education and empowerment by providing answers to medical questions and translating complex medical information into more accessible language[4,13]. Thereby, LLMs may promote personalized medicine and broaden access to medical knowledge, empowering patients to actively participate in their healthcare decisions.

To the best of our knowledge, there has been no evaluation of existing research to understand the scope of applications and identify limitations that may currently limit the successful integration of LLMs into clinical practice. This systematic review aims to analyze and synthesize the literature on LLMs in patient care, providing a systematic overview of 1) current applications and 2) challenges and limitations, with the purpose of establishing a foundational framework and taxonomy for the implementation and evaluation of LLMs in healthcare settings.

Methods

This systematic review was pre-registered in the International Prospective Register of Systematic Reviews (PROSPERO) under the identifier CRD42024504542 before the start of the initial screening and was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyzes (PRISMA) guidelines (see checklist in the Supplementary Dataset file 1)[14,15].

Eligibility criteria

We searched 5 databases, including the Web of Science, PubMed, Embase/Embase Classic, American for Computing Machinery (ACM) Digital Library, and Institute of Electrical and Electronics Engineers (IEEE) Xplore as of January 25, 2024, to identify qualitative, quantitative, and mixed methods studies published between January 1, 2022, and December 31, 2023, that examined the use of LLMs for patient care. LLMs for patient care were defined as any artificial neural network that follows a transformer architecture and can be used to generate and translate text and other content or perform other natural language processing tasks for the purpose of disease management and support (i.e., prevention, preclinical management, diagnosis, treatment, or prognosis) that could be directly directed to or used by patients. Articles had to be available in English and contain sufficient data for thematic synthesis (e.g., conference abstracts that did not provide sufficient information on study results were excluded). Given the recent surge in publications on LLMs such as ChatGPT, we allowed for the inclusion of preprints if no corresponding peer-reviewed article was available. Duplicate reports of the same study, non-human studies, and articles limited to technology development/performance evaluation, pharmacy, human genetics, epidemiology, psychology, psychosocial support, or behavioral assessment were excluded.

Screening and data extraction

Initially, we conducted a preliminary search on PubMed and Google Scholar to define relevant search terms. The final search strategy included terms for LLMs, generative AI, and their applications in medicine, health services, clinical practices, medical treatments, and patient care (as detailed by database in the Supplementary Methods). After importing the bibliographic data into Rayyan and removing duplicates, LH and CR conducted an independent blind review of each article's title and abstract[16]. Any article flagged as potentially eligible by either reviewer proceeded to the full-text evaluation stage. For this stage, LH and CR used a custom data extraction form created in Google Forms (available online)[17] to collect all relevant data independently from the studies that met the inclusion criteria. Quality assessment was also performed independently for each article within this data extraction form, using the Mixed Methods Appraisal Tool (MMAT) 2018[18]. Disagreements at any stage of the review were resolved through discussion with the author FB. In cases of studies with incomplete data, we have tried to contact the corresponding authors for clarification or additional information.

Data analysis

Due to the diversity of investigated outcomes and study designs we sought to include, a meta-analysis was not practical. Instead, a data-driven convergent synthesis approach was selected for thematic syntheses of LLM applications and limitations in patient care[19]. Following Thomas and Harden, FB coded each study's numerical and textual data in Dedoose using free line-by-line coding[20,21]. Initial codes were then systematically categorized into descriptive and subsequently into analytic themes, incorporating new codes for emerging concepts within a hierarchical tree structure. Upon completion of the codebook, FB and LH reviewed each study to ensure consistent application of codes. Discrepancies were resolved through discussion with the author KKB, and the final codebook and analytical themes were discussed and refined in consultation with all contributing authors.

Results

Screening results

Of the 4349 reports identified, 2991 underwent initial screening, and 126 were deemed suitable for potential inclusion and underwent full-text screening. Two articles could not be retrieved because the authors or the corresponding title and abstract could not be identified online. Following full-text screening, 35 articles were excluded, and 89 articles were included in the final review. Most studies were excluded because they targeted the wrong discipline ($n = 10/35$, 28.6%) or population ($n = 7/35$, 20%) or were not original research ($n = 8/35$, 22.9%) (see Supplementary Dataset file 2). For example, we evaluated a study that focused on classifying physician notes to identify patients without active bleeding who were appropriate candidates for thromboembolism prophylaxis[22]. Although the classification tasks may lead to patient treatment, the primary outcome was informing clinicians rather than directly forwarding this information to patients. We also reviewed a study assessing the accuracy and completeness of several LLMs when answering Methotrexate-related questions[23]. This study was excluded because it focused solely on the pharmacological treatment of rheumatic disease. For a detailed breakdown of the inclusion and exclusion process at each stage, please refer to the PRISMA flowchart in Fig. 1.

Characteristics of included studies

Supplementary Dataset file 3 summarizes the characteristics of the analyzed studies, including their setting, results, and conclusions. One study ($n = 1/89$, 1.1%) was published in 2022[24], 84 ($n = 84/89$, 94.4%) in 2023[13,25–107], and 4 ($n = 4/89$, 4.5%) in 2024[108–111] (all of which were peer-reviewed publications of preprints published in 2023). Most studies were quantitative non-randomized ($n = 84/89$, 94.4%)[13,25–27,29–101,103,104,106,107,109–111], 4 ($n = 4/89$, 4.5%)[28,102,105,108] had a qualitative study design, and one ($n = 1/89$, 1.1%)[24] was quantitative randomized according to the MMAT 2018 criteria. However, the LLM outputs were often first analyzed quantitatively but followed by a qualitative analysis of certain responses. Therefore, if the primary outcome was quantitative, we considered the study design to be quantitative rather than mixed methods, resulting in the inclusion of zero mixed methods studies. The quality of the included studies was mixed (see Supplementary Dataset file 4). The authors were primarily affiliated with institutions in the United States ($n = 47$ of 122 different countries identified per publication, 38.5%), followed by Germany ($n = 11/122$, 9%), Turkey ($n = 7/122$, 5.7%),

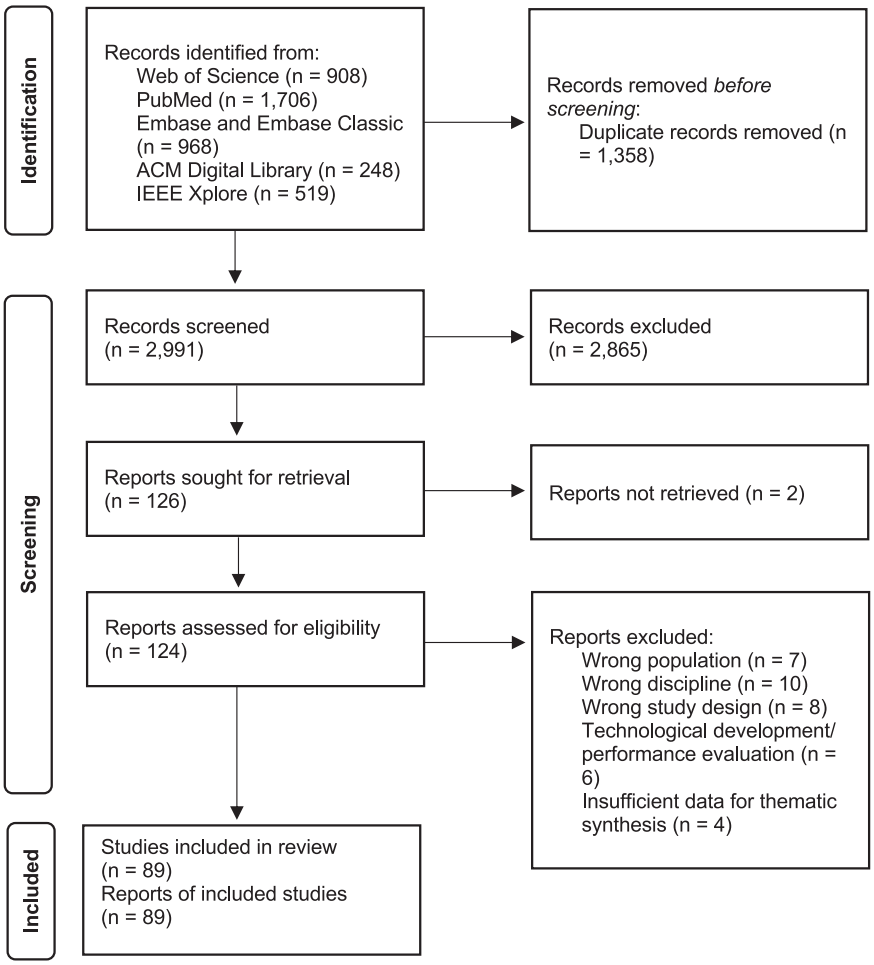

Fig. 1 | Preferred reporting items for systematic reviews and meta-analyzes (PRISMA) flow diagram. A total of 4349 reports were identified from Web of Science, PubMed, Embase/Embase Classic, ACM Digital Library, and IEEE Xplore. After excluding 1358 duplicates, 2991 underwent initial screening and 126 were deemed suitable for potential inclusion and underwent full-text screening. Two articles could not be retrieved because the authors or the corresponding title and abstract could not be identified online. After full text screening, 35 articles were excluded and 89 articles were included in the final review.

the United Kingdom ($n = 6/122$, 4.9%), China/Australia/Italy ($n = 5/122$, 4.1%, respectively), and 24 ($n = 36/122$, 29.5%) other countries. Most studies examined one or more applications based on the GPT-3.5 architecture ($n = 66$ of 124 different LLMs examined per study, 53.2%)[13,26–29,31–34,36–40,42–49,52–54,56–61,63,65–67,71,72,74,75,77,78,81–89,91,92,94,95,97–100,102–104,106–109,111], followed by GPT-4 ($n = 33/124$, 26.6%)[13,25,27,29,30,34–36,41,43,50,51,54,55,58,61,64,68–70,74,76,79–81,83,87,89,90,93,96,98,99,101,105], Bard ($n = 10/124$, 8.1%; now known as Gemini)[33,48,49,55,73,74,80,87,94,99], Bing Chat ($n = 7/124$, 5.7%; now Microsoft Copilot)[49,51,55,73,94,99,110], and other applications based on Bidirectional Encoder Representations from Transformers (BERT; $n = 4/124$, 3.2%)[13,83,84], Large Language Model Meta-AI (LLaMA; $n = 3/124$, 2.4%)[55], or Claude by Anthropic ($n = 1/124$, 0.8%)[55]. The majority of applications were primarily targeted at patients ($n = 64$ of 89 included studies, 73%)[24,25,29,32,34–39,41–43,45–48,52–54,56–60,62,63,65,66,68–71,73–75,77–80,85–95,97,99,100,102–111] or both patients and caregivers ($n = 25/89$, 27%)[13,26–28,30,31,33,40,44,49–51,55,61,64,67,72,76,81–84,96,98,101]. Information about conflicts of interest and funding was not explicitly stated in 23 ($n = 23/89$, 25.8%) studies, while 48 ($n = 48/89$, 53.9%) reported that there were no conflicts of interest or funding. A total of 18 ($n = 18/89$, 20.2%) studies reported the presence of conflicts of interest and funding[13,24,38,40,54,58,59,67,69–71,74,80,84,96,103,105,111]. Most studies did not report information about the institutional review board (IRB) approval ($n = 55/89$, 61.8%) or deemed IRB approval unnecessary ($n = 28/89$, 31.5%). Six studies obtained IRB approval ($n = 6/89$, 6.7%)[52,82,84–86,92].

Applications of large language models

An overview of the presence of codes for each study is provided in the Supplementary Dataset file 3. The majority of articles investigated the use and feasibility of LLMs as medical chatbots ($n = 84/89$, 94.4%)[13,24–62,64–66,68,69,71–96,98–111], while fewer reports additionally or exclusively focused on the generation of patient information ($n = 18/89$, 20.2%)[24,31,43,48,49,57,59,62,67,79,88–91,97,102,106,107], including clinical documentation such as informed consent forms ($n = 5/89$, 5.6%)[43,67,91,97,102] and discharge instructions ($n = 1/89$, 1.1%)[31], or translation/summarization tasks of medical texts ($n = 5/89$, 5.6%)[24,49,57,79,89], creation of patient education materials ($n = 5/89$, 5.6%)[48,62,90,106,107], and simplification of radiology reports ($n = 2/89$, 2.3%)[59,88]. Most reports evaluated LLMs in English ($n = 88/89$, 98.9%)[13,24–103,105–111], followed by Arabic ($n = 2/84$, 2.3%)[32,104], Mandarin ($n = 2/84$, 2.3%)[36,75], and Korean or Spanish ($n = 1/89$, 1.1%, respectively)[75]. The top-five specialties studied were ophthalmology ($n = 10/89$, 11.2%)[37,40,48,51,65,74,97,98,100,101], gastroenterology ($n = 9/89$, 10.1%)[25,32,34,36,39,61,62,72,96], head and neck surgery/otolaryngology ($n = 8/89$, 9%)[35,42,56,64,66,76,78,79], and radiology[59,70,88–90,110] or plastic surgery[45,47,49,102,107,108] ($n = 6/89$, 6.7%, respectively). A schematic illustration of the identified concepts of LLM applications in patient care is shown in Fig. 2.

Limitations of large language models

The thematic synthesis of limitations resulted in two main concepts: one related to design limitations and one related to output. Figure 3 illustrates the hierarchical tree structure and quantity of the codes derived from the thematic synthesis of limitations. Supplementary Dataset file 5 provides an overview of the taxonomy of all identified limitation concepts, including their description and examples.

Design limitations. In terms of design limitations, many authors noted the limitation that LLMs are not optimized for medical use ($n = 46/89$, 51.7%)[13,26,28,34,35,37–39,46,49,50,54–59,61,62,65,66,68,70,71,79–81,83–85,88,91,93–98,100–107,109], including implicit knowledge/lack of clinical context ($n = 13/89$, 14.6%)[28,39,46,66,71,79,81,83–85,98,103], limitations in clinical reasoning ($n = 7/89$,

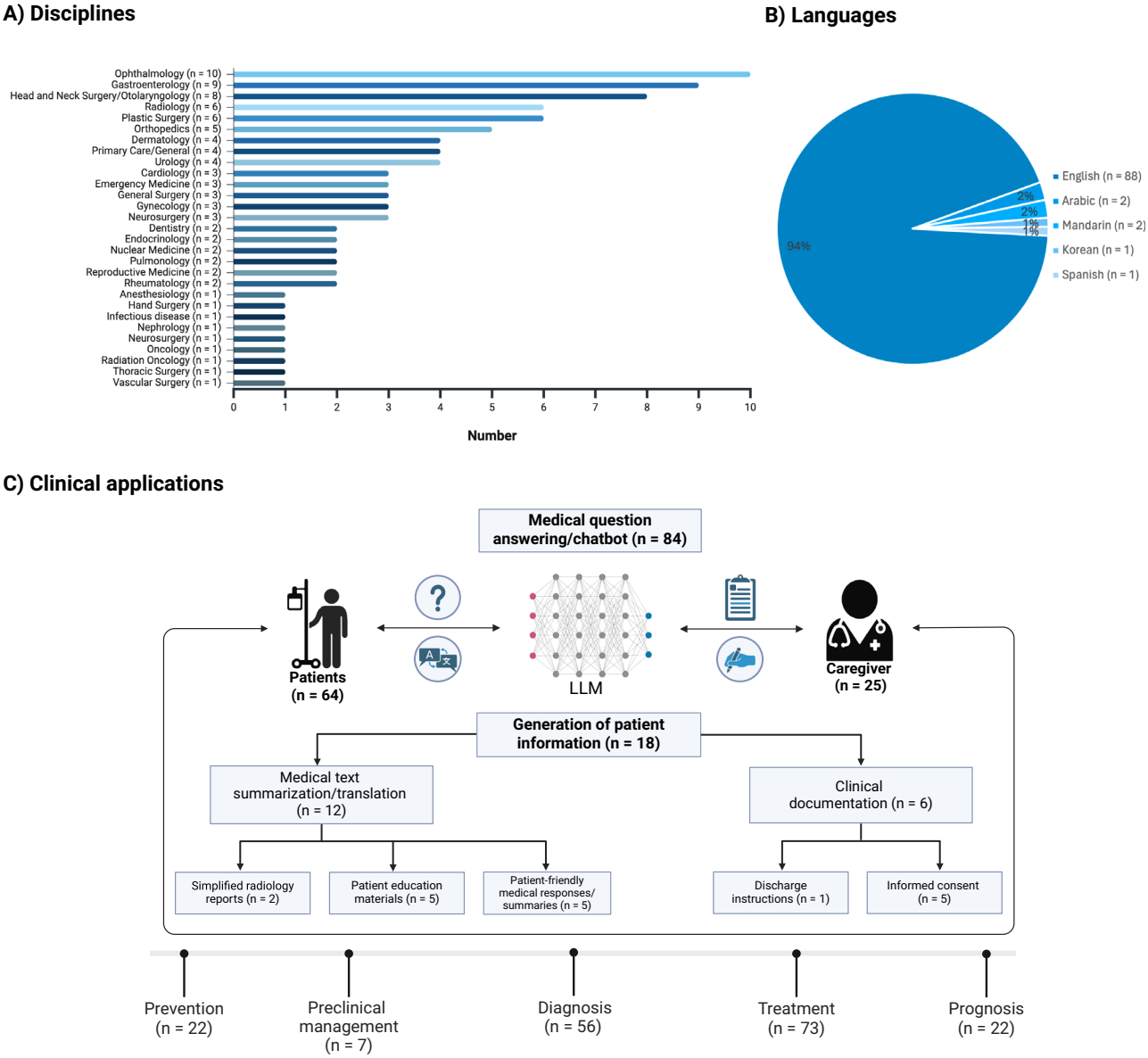

Fig. 2 | Schematic illustration of the identified disciplines, languages, and clinical concepts of large language models (LLMs) applications in patient care. A Column plot showing the distribution of medical specialties in which LLMs have been tested for patient care. **B** Pie chart illustrating the distribution of languages in which LLMs have been tested. **C** Schematic representation of the concepts identified for the application of LLMs in patient care.

7.9%)[55,84,95,102–105], limitations in medical image processing/production ($n = 5/89$, 5.6%)[37,55,91,106,107], and misunderstanding of medical information and terms by the model ($n = 7/89$, 7.9%)[28,38,39,59,62,65,97]. In addition, data-related limitations were identified, including limited access to data on the internet ($n = 22/89$, 24.7%)[38,39,41,43,54–57,59,60,64,76,79,82–84,88,91,94,96,104,109], the undisclosed origin of training data ($n = 36/89$, 40.5%)[25,26,29,30,32,34,36,37,40,46,47,50,51,53–60,64,65,70,71,76,82,83,91,94–96,101,105,109], limitations in providing, evaluating, and validating references ($n = 20/89$, 22.5%)[45,49,54–57,65,71,73,76,80,83,85,91,94,96,98,101,103,105], and storage/processing of sensitive health information ($n = 8/89$, 9%)[13,34,46,55,62,76,83,109]. Further second-order concepts included black-box algorithms, i.e., non-explainable AI ($n = 12/89$, 13.5%)[27,36,55,57,65,73,76,83,91,94,103,105], limited engagement and dialog capabilities ($n = 10/89$, 11.2%)[13,27,28,37,38,51,56,66,95,103], and the inability of self-validation and correction ($n = 4/89$, 4.5%)[61,73,74,107].

Output limitations. The evaluation of limitations in output data yielded 7 second-order codes concerning the non-reproducibility ($n = 38/89$, 42.7%)[28,29,34,38,39,41,43,45,46,49,54–61,64,65,71–73,76,80,82,83,85,90,91,94,96,98,99,101,103–105], non-comprehensiveness ($n = 78/89$, 87.6%)[13,25,26,28–30,32–44,46,48–62,64,65,67–79,81–98,100,102–107,109–111], incorrectness ($n = 78/89$, 87.6%)[13,25–44,46,49–52,54–62,64–66,69–79,81–85,87–107,109–111], (un-)safeness ($n = 39/89$, 43.8%)[28,30,35,37,39,40,42–44,46,50,51,57–60,62,64,65,69,70,73,74,76,78–80,82,84,85,91,94,95,98–100,105,106,109], bias ($n = 6/89$, 6.7%)[26,32,34,36,66,103], and the dependence of the quality of output on the prompt-/input provided ($n = 27/89$, 30.3%)[26–28,34,38,41,44,46,51,52,56,68–72,74,76,78,79,81–83,90,94,95,100,101] or the environment ($n = 16/89$, 18%)[13,34,46,49–51,54,58,60,72,73,88,90,93,97,109].

Non-reproducibility. For non-reproducibility, key concepts included the non-deterministic nature of the output, e.g., due to inconsistent results across multiple iterations ($n = 34/89$, 38.2%)[28,29,34,38,39,41,43,46,58–61,72,76,82,90,94,98,99,101,103,104] and the inability to provide reliable references ($n = 20/89$, 22.5%)[45,49,54–57,65,71,73,76,80,83,85,91,94,96,98,101,103,105].

Non-comprehensiveness. Non-comprehensiveness included nine concepts related to generic/non-personalized output ($n = 34/89$,

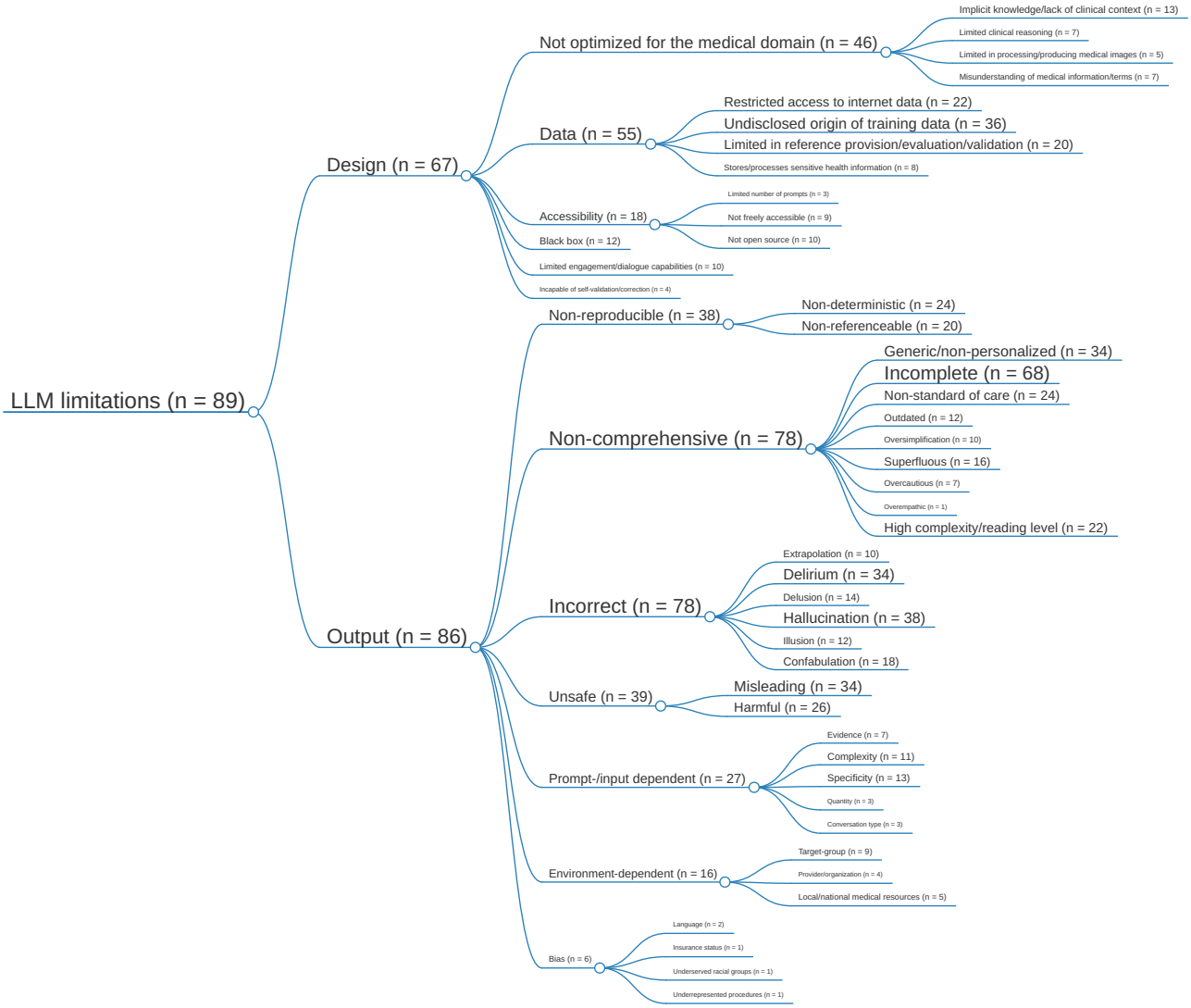

Fig. 3 | Illustration of the hierarchical tree structure for the thematic synthesis of large language model (LLM) limitations in patient care, including the presence of codes for each concept. The font size of each concept is shown in proportion to its frequency in the studies analyzed. Our analysis delineates two primary domains of LLM limitations: design and output. Design limitations included 6 second-order and 12 third-order codes, while output limitations included 9 second-order and 32 third-order codes.

38.2%)[13,28,30,34,37,38,41,43,49,51,56,57,59,61,65,70,77,79,81,84–86,90,94,95,100,102–107,110], incompleteness of output ($n = 68/89$, 76.4%)[13,25,26,28–30,32,34–39,41–44,46,49–52,55–62,64,65,67–69,72–77,79,81–86,89–98,100,102–107,109–111], provision of information that is not standard of care ($n = 24/89$, 27%)[28,40,43,46,49,50,54,57,58,65,69,72,73,77,78,81,85,91,94,98,100,103,107,111] and/or outdated ($n = 12/89$, 13.5%)[13,25,32,34,38,41,43,44,49,54,83,84], and production of oversimplified ($n = 10/89$, 11.2%)[38,46,49,54,59,79,84,85,103], superfluous ($n = 16/89$, 18%)[13,28,34,38,46,62,72,79,86,90,94,97,100,106,107], overcautious ($n = 7/89$, 7.9%)[13,28,37,51,70,103,110], overempathic ($n = 1/89$, 1.1%)[13], or output with inappropriate complexity/reading level for patients ($n = 22/89$, 24.7%)[13,34,42,48,50,51,53,55,56,67,71,78,79,85,87,88,90,93,106,107,109,110].

Incorrectness. For incorrectness, we identified 6 key concepts. Some of the incorrect information could be attributed to what is commonly known as hallucination ($n = 38/89$, 42.7%)[25,28,32,33,35–38,40–44,49–51,57–60,65,73,74,76,77,81,83,85,91,94,96–98,100,103,106,107,109], i.e., the creation of entirely fictitious or false information that has no basis in the input provided or in reality (e.g., "You may be asked to avoid eating or drinking for a few hours before the scan" for a bone scan). Other instances of misinformation were more appropriately classified under alternative concepts of the original psychiatric analogy, as described in detail by Currie et al.[43,112,113]. These include illusion ($n = 12/89$, 13.5%)[28,36,38,43,57,59,77,78,85,88,94,105], which is characterized by the generation of deceptive perceptions or the distortion of information by conflating similar but separate concepts (e.g., suggesting that MRI-type sounds might be experienced during standard nuclear medicine imaging), delirium ($n = 34/89$, 38.2%)[13,26,28,30,37,43,50,58,59,61,65,70,72–75,77,79,81–85,90–92,94,95,98,102,103,107,109,110], which indicates significant gaps in vital information, resulting in a fragmented or confused understanding of a subject (e.g., omission of crucial information about caffeine cessation for stress myocardial perfusion scans), extrapolation ($n = 11/89$, 12.4%)[43,59,65,78,81,91,94,106,107,110], which involves applying general knowledge or patterns to specific situations where they are inapplicable (e.g., advice about injection-site discomfort that is more typical of CT contrast administration), delusion ($n = 14/89$, 15.7%)[28,30,43,50,59,65,69,73,74,78,81,94,103,111], a fixed, false belief despite contradictory evidence (e.g., inaccurate waiting times for the thyroid scan), and confabulation ($n = 18/89$, 20.2%)[25,28,36–38,40,46,59,62,65,71,77–79,94,103,107], i.e., filling in memory or knowledge gaps with plausible but invented information (e.g., "You should drink plenty of fluids to help flush the radioactive material from your body" for a biliary system–excreted radiopharmaceutical).

Safety and bias. Many studies rated the generated output as unsafe, including misleading ($n = 34/89$, 38.2%)[28,30,35,43,44,46,50,51,57–60,62,64,65,69,73,74,76,

[78–80,82,84,85,94,95,98–100,105,106,109] or even harmful content ($n = 26/89$, 29.2%)[28,30,37,39,40,42,43,50,51,58–60,70,73,74,76,79,84,85,91,94,95,98–100,109].

A minority of reports identified biases in the output, which were related to language ($n = 2/89$, 2.3%)[32,36], insurance status[103], underserved racial groups[26], or underrepresented procedures[34] ($n = 1/89$, 1.1%, each).

Dependence on input and environment. Many authors suggested that performance was related to the prompting/input provided or the environment, i.e., depending on the evidence ($n = 7/89$, 7.9%)[52,68,69,71,81,82,95], complexity ($n = 11/89$, 12.4%)[28,34,44,46,70,74,76,79,94,102], specificity ($n = 13/89$, 14.6%)[27,38,41,56,70,72,74,76,78,81,95,100,101], quantity ($n = 3/89$, 3.4%)[26,52,74] of the input, type of conversation ($n = 3/89$, 3.4%)[27,51,90], or the appropriateness of the output related to the target group ($n = 9/89$, 10.1%)[46,49,51,54,72,90,93,97,109], provider/organization ($n = 4/89$, 4.5%)[13,50,60,88], and local/national medical resources ($n = 5/89$, 5.6%)[34,50,58,60,73].

Discussion

In this systematic review, we synthesized the current applications and limitations of LLMs in patient care, incorporating a broad analysis across 29 medical specialties and highlighting key limitations in LLM design and output, providing a comprehensive framework and taxonomy for describing and categorizing limitations that may arise when using LLMs in healthcare settings.

Most articles examined the use of LLMs based on the GPT-3.5 or GPT-4 architecture for answering medical questions, followed by the generation of patient information, including medical text summarization or translation and clinical documentation. The conceptual synthesis of LLM limitations revealed two key concepts: the first related to design, including 6 second-order and 12 third-order codes, and the second related to output, including 9 second-order and 32 third-order codes. By systematically categorizing the limitations of LLMs in clinical settings, our taxonomy aims to provide healthcare professionals and developers with a framework for assessing potential risks associated with the use of LLMs in patient care. In addition, our work highlights key areas for improvement in the development of LLMs and aims to enable clinicians to make more informed decisions by understanding the limitations inherent in the design and output, thereby supporting the establishment of best practices for LLM use in clinical settings.

Although many LLMs have been developed specifically for the biomedical domain in recent years, we found that ChatGPT has been a disruptor in the medical literature on LLMs, with GPT-3.5 and GPT-4 accounting for almost 80% of the LLMs examined in this systematic review. While it was not possible to conduct a meta-analysis of the performance on medical tasks, many authors provided a positive outlook towards the integration of LLMs into clinical practice. However, we have conceptualized several key limitations in the design and output of LLMs, some of the most prevalent in our systematic review are briefly discussed in the following paragraphs.

The majority of studies ($n = 55/89$) reported limitations that were conceptualized as related to the underlying data of the LLMs studied. Especially the use of proprietary models such as ChatGPT in the biomedical field was a concern in many of the studies analyzed, mainly because of the lack of training data transparency (third-order code: undisclosed origin of training data). In practice, it is widely recognized that limited access to the underlying algorithms, training data, and data processing and storage mechanisms of LLMs is a significant barrier to their application in healthcare[114]. This opacity makes it difficult for healthcare professionals to fully understand how these models function, assess their reliability, or ensure compliance with local medical standards and regulations. Consequently, the use of such models in healthcare settings can be problematic, and the need to recognize and correct potential limitations in the outputs of such models is paramount.

Moreover, integrating proprietary models into clinical practice introduces a vulnerability to performance changes that occur with model updates[115]. As these models are updated by their developers, functionalities that healthcare providers rely on may be altered or broken, potentially leading to harmful outcomes for patients, which was also conceptualized in our study under output limitations (second-order code: unsafe; third-order codes: misleading/harmful). This unpredictability is a serious concern in the biomedical field, where consistency and reliability are crucial. Notably, the unpredictability of LLMs was another concept of output limitations in our systematic review (second-order code: non-reproducible; third-order codes: non-deterministic/non-referenceable).

As a result, open-source models such as BioMistral may offer a viable solution[6]. Such open source models not only offer more transparency, as their algorithms and training data are accessible but can also be adapted locally. However, given the limited number of articles on open-source LLMs in our review, we strongly encourage future studies investigating the applicability of open-source LLMs in patient care.

About half of the studies analyzed reported limitations related to LLMs not being optimized for the medical domain. One possible solution to this limitation may be to provide medical knowledge during inference using RAG[116]. However, even when trained for general purposes, ChatGPT has previously been shown to pass the United States Medical Licensing Examination (USMLE), the German State Examination in Medicine, or even a radiology board-style examination without images[117–120]. Although outperformed on specific tasks by specialized medical LLMs, such as Google's MedPaLM-2, this suggests that general-purpose LLMs can comprehend complex medical literature and case scenarios to a degree that meets professional standards[121]. Furthermore, given the large amounts of data on which proprietary models such as ChatGPT are trained, it is not unlikely that they have been exposed to more medical data overall than smaller specialized models despite being generalist models. Notably, a recent study even suggested that fine-tuning LLMs on biomedical data does not improve performance compared to their general-purpose counterparts[122].

It should also be noted that passing these exams does not equate to the practical competence required of a healthcare provider, which was also a limitation identified in our review (third-order codes: implicit knowledge/lack of clinical context; limited clinical reasoning; misunderstanding of medical information/terms; limited in processing/producing medical images)[123]. In addition, reliance on exam-based assessments carries a significant risk of bias. For example, if the exam questions or similar variants are publicly available and, thus, may be present in the training data, the LLM does not demonstrate any knowledge outside of training data memorization[124]. In fact, these types of tests can be misleading in estimating the model's true abilities in terms of comprehension or analytical skills.

The non-reproducibility of LLM output, as conceptualized in 38 studies, highlights key challenges in ensuring consistency and determinism in LLM-generated results. One major issue is the inherent stochasticity in the models' architecture, particularly in transformer-based models, which utilize probabilistic techniques during inference (e.g., beam search or temperature sampling)[125]. This non-determinism can lead to different outputs for the same input, making it difficult to replicate results exactly across different instances or even across models with identical training data. Further external factors contributing to non-reproducibility, such as variations in hardware, software versions, or context windows, complicate the assurance of reproducibility[126]. As the reproducibility of results is a central principle in medical practice, our concepts highlight the need for more standardized protocols, improved documentation of model configurations, the examination of non-determinism for evaluation purposes, and further research on how robust results can be achieved before implementing LLMs in real-world clinical practice. Interestingly, Ouyang et al. reported that only a minority of studies take non-determinism into account in their experimental evaluation when using ChatGPT for code generation, suggesting that this limitation is also prevalent and overlooked in other domains of LLM use[125].

The concept of non-comprehensiveness was prevalent in almost 90% of the studies analyzed ($n = 78/89$). For this concept, the majority of third-order codes were related to LLM outputs that were incomplete. This issue is particularly significant when considering the application of LLMs in medical tasks such as clinical decision support or diagnosis, where incomplete or partial results can have serious consequences. In clinical practice, missing key information could lead to suboptimal patient outcomes, incorrect diagnoses, or improper treatment recommendations. For instance, an incomplete therapy suggestion could render the entire treatment plan

insufficient, potentially resulting in harm to the patient. Given the potential of using LLMs in medical decision-making, these limitations underscore the necessity for expert supervision and validation of LLM outputs depending on their application. While LLMs used as chatbots for general patient inquiries may not require consistent human oversight, using LLMs for treatment advice would require consistent validation to ensure that incomplete information does not lead to adverse outcomes. Depending on their application, the same problem arises when the LLM generates generic or non-personalized information, which was another third-order code identified. The generation of content with high complexity and an inappropriate reading level, which was above the American Medical Association (AMA) recommended 6th-grade reading level in almost all of the 22 studies that analyzed the complexity level of the output, may further limit its usefulness for patient information[127]. Again, the best solution to the lack of comprehensiveness in clinical practice so far seems to be human oversight.

Incorrectness, alongside non-comprehensiveness (as above), was the most common second-order code, identified in about 90% of studies (n = 78/89). In our conceptual synthesis of incorrect results, we followed the taxonomy of Currie et al. to classify incorrect outputs more precisely into illusions, delusions, delirium, confabulation, and extrapolation, thus proposing a framework for a more precise and structured error classification to improve the characterization of incorrect outputs and enabling more detailed performance comparisons with other research[43,112,113].

Many studies currently refer to all non-factual LLM results as "hallucinations." However, this generalization fails to capture the complexity of errors when considering the original psychiatric analogy. Simply classifying errors as hallucinations restricts their description to invented information, overlooking errors that, for example, omit critical information and leading to fragmented or confused understanding (third-order code: delirium). Notably, the third-order code "delirium" was observed in nearly as many studies as the third-order code "hallucination." However, a non-detailed classification of incorrect results can affect not only the comparability of research findings but also has implications for clinical practice. While hallucinations (for example, fabricating instructions like "You may need to fast before a bone scan.") may not always have serious consequences, errors classified as delirium—such as omitting crucial details like caffeine cessation before a stress myocardial perfusion scan—would always result in undesired outcomes (in the here presented example, most likely in repeating or postponing the examination). As a result, our review advocates for a more detailed classification of incorrect results in order to increase the qualitative comparability of incorrect LLM outputs and, ultimately, the relevance and implications of these results for clinical practice.

The conceptualization of unsafeness in 39 of 89 studies presents a significant concern when considering the integration of LLMs into medical practice. In the field of medicine, any tool or intervention that could lead to misleading or harmful outcomes must be critically assessed, as the potential for patient harm is high[128,129]. Such tools are generally only accepted when the benefits clearly outweigh the risks, and even then, informed consent from the affected individual is essential[130]. While informed consent might ensure that patients understand the risks involved and are able to make an educated decision about their care, which could be obtained, for example, in the form of a disclaimer before using the LLM, studies suggest that even when obtaining informed consent the patient understanding increases not significantly[131]. In the case of disclaimers, there might also be the risk that these are accepted without proper reading or understanding[132]. The practicality of informed consent once LLMs are deeply integrated into clinical workflows also remains an issue, as it is when patients no longer have the ability to opt out, such as in the case of serious illness. In any case, the finding that nearly half of the studies reported limitations related to LLM unsafeness suggests that LLMs are not yet reliable enough for autonomous medical use, and there is a critical need for safety measures and regulatory and human oversight to prevent adverse consequences in medical contexts[133].

Further second-order concepts suggested that the output is influenced by the input or environment in which it is expressed. In fact, LLMs can be highly dependent on the quality and specificity of input, making their output prone to errors when faced with vague or incomplete information[134–136]. Again, this poses significant risks in patient care, where incorrect outputs can lead to adverse outcomes, such as inappropriate triage or treatment. For example, in our review, eleven studies reported a decrease in performance with increasing complexity of the input, which can have implications in clinical practice, such as failing to consider multifactorial medical issues like comorbidities, thus compromising the quality of care for those patients.

We found that the environment also influences the appropriateness of LLM outputs in medical settings. Models may recommend treatments that are inappropriate for certain patient populations, such as offering adult care protocols for pediatric patients or suggesting therapies that are not available in certain regions. This also raises ethical concerns, particularly in resource-constrained settings, where making inappropriate or inaccessible recommendations may reinforce existing inequalities and lead to uncomfortable situations for both the healthcare provider and the patient[137]. One solution may be to provide adequate training to LLM users, or in our scenario, to patients, on how to present input to the model to achieve the best results. Another solution is to train or fine-tune the model to the environment in which it will be used. For example, if the LLM is trained on the standard operating procedure for handling patients with major adverse cardiovascular events in a particular hospital, it is more likely to recommend the adequate procedure in this setting than when it is trained on worldwide data from an unknown time frame, where there is a chance that it will suggest non-standard care that may only be relevant in other countries where most of the training data is coming from, or even provides outdated information (which is another third-order code that was conceptualized under non-comprehensiveness) if it is trained on data that is not current.

Ultimately, only six studies have identified biases in their results, for example, reflecting the unequal representation of certain content or the biases inherent in human-generated text in the training data[138]. Here, we conceptualized the results of studies that identified bias in their analysis and not only mentioned bias as a theoretical limitation. Thus, these results may indicate that the implemented safeguards are effective. On the other hand, identifying bias was not the primary outcome of most studies, and not much is known about the technology and developer policies of proprietary LLMs. Moreover, previous work has shown that automated jailbreak generation is possible across various commercial LLM chatbots[139]. In the end, LLMs are trained on large datasets that inevitably contain biases—such as gender, racial, or cultural biases—embedded in the text[140]. These biases can be amplified or reflected by the models, leading to unfair or harmful outputs. Despite the use of various mitigation techniques, such as debiasing algorithms, curating balanced datasets, or incorporating fairness-focused training objectives, eliminating bias entirely is a persistent challenge[141–143]. This is because LLMs learn patterns from their training data, and human biases are inherently present in much of the data they consume. Moreover, the biases introduced or reinforced by LLM are not always obvious, making them more difficult to detect and correct, which may have contributed to the comparatively low number of studies that reported any bias in their results. Notably, subtle biases, such as those related to linguistic connotations, regional dialects, or implicit associations, can be especially insidious and difficult to eliminate through technical safeguards[144]. Therefore, the results of our review may encourage future studies to more explicitly examine the biases inherent in LLMs when used for medical tasks and how such biases could be mitigated.

Our findings raise a key question when applying LLMs to the medical domain: how can we entrust our patients to LLMs if they are neither reliable nor transparent? Given that models like ChatGPT are already publicly accessible and widely used, patients may already refer to them for medical questions in much the same way they use Google Search, making concerns about their early adoption somewhat academic[145].

In addition to the advances in the development of LLMs and the focus on open source, adopting appropriate security measures to prevent the identified LLM limitations in clinical practice out-of-the-box will become increasingly important. For example, strategies to ensure LLM security and privacy can include continuous monitoring for new vulnerabilities, implementing input validation, conducting regular audits of training data, and

using secure data pipelines[146]. Additionally, data anonymization, encryption, access controls, and regular security updates are essential to prevent data leakage, model theft, and privacy breaches.

Moreover, expert oversight of the final LLM output could mitigate any remaining risks in the last instance, ensuring that erroneous or inappropriate suggestions are identified and corrected before they can impact patient care. Recently, efforts have been made in this direction by adopting the widely recognized Physician Documentation Quality Instrument (PDQI-9) for the assessment of AI transcripts and clinical summaries[147]. However, whether ongoing human oversight and validation of LLM-generated content is feasible and can reduce the likelihood of adverse outcomes remains the subject of further research at this early stage of LLM deployment in healthcare.

Another important factor for the successful clinical implementation of LLMs in patient care could be patient acceptance, which was not assessed in any of the studies analyzed. The growing use of LLMs in healthcare might be perceived as a reduction in the interpersonal relationship between healthcare professionals and patients, potentially leading to a sense of dehumanization in medicine[148]. Therefore, to promote a positive reception of AI tools among patients, incorporating their perspectives already during the AI development and implementation process could be key[149]. Eventually, patient perspectives are already considered in AI regulatory frameworks, such as in the European Union AI Act, which came into force in August 2024[150]. The associated challenges faced by generative AI and LLM, for example, in terms of training data transparency and validation of non-deterministic output, will show which approaches the companies will take to bring these models into compliance with the law[151]. How the notified bodies interpret and enforce the law in practice will likely be decisive for the further development of LLMs in the biomedical sector.

Our study has limitations. First, our review focused on LLM applications and limitations in patient care, thus excluding research directed at clinicians only. Future studies may extend our synthesis approach to LLM applications that explicitly focus on healthcare professionals. Second, while it was not possible to conduct a meta-analysis of LLM performance due to the different study designs and evaluation methods used, this will be an important area for future work as the field of LLM research in clinical settings continues to evolve. Third, there is a risk that potentially eligible studies were not included in our analysis if they were not present in the 5 databases reviewed or were not available in English. However, we screened nearly 3,000 articles in total and systematically analyzed 89 articles, providing a comprehensive overview of the current state of LLMs in patient care, even if some articles could have been missed. With our chosen cut-off date of January 2022, there is also a risk of missing relevant publications on predecessor LLM models, such as GPT-3, which was introduced in 2020. However, as our review focused on current LLM applications and limitations, it seemed most beneficial to include only recent publications from the last two years on the most advanced models, especially when considering that ChatGPT was first made available in November 2022. Finally, the rapid development and advancement of LLMs make it difficult to keep this systematic review up to date. For example, Gemini 1.5 Pro was published in February 2024, and corresponding articles are not included in this review, which synthesized articles from 2022 to 2023. This also has implications for our introduced taxonomy of LLM limitations, as new limitations may emerge as models evolve, and previous limitations may become less relevant or even obsolete. For example, our taxonomy identifies "limited access to internet data" as a limitation; however, with the introduction of web browsing capabilities for GPT-4 in May 2023, this particular limitation no longer applies to that model. Given these ongoing developments, we strongly encourage future studies to test, update, and extend our taxonomy to ensure that it remains a relevant tool for categorizing LLM limitations in clinical and other high-stakes applications.

Data availability

All data generated or analyzed during this study, including source data, are included in this published article and its supplementary information files. To provide a more intuitive overview and allow readers to filter through the collected study data and codes, we have provided the Supplementary Dataset files in the form of Excel spreadsheets.

Received: 11 March 2024; Accepted: 17 December 2024;
Published online: 21 January 2025

References

1. Milmo, D. *ChatGPT reaches 100 million users two months after launch*, https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app (February 2, 2023).
2. OpenAI. GPT-4 Technical Report. arXiv:2303.08774. https://ui.adsabs.harvard.edu/abs/2023arXiv230308774O (2023).
3. Zhao, W. X. et al. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
4. Clusmann, J. et al. The future landscape of large language models in medicine. *Communications Medicine* **3**, 141 (2023).
5. Chen, Z. et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079* (2023).
6. Labrak, Y. et al. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv preprint arXiv:2402.10373 (2024).
7. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. *arXiv preprint arXiv:2402.13178* (2024).
8. Yang, X. et al. A large language model for electronic health records. *npj Digital Medicine* **5**, 194 (2022).
9. Tian, S. et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics* **25**. https://doi.org/10.1093/bib/bbad493 (2024)
10. Adams, L. C. et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* **307**, e230725 (2023).
11. McDuff, D. et al. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164* (2023).
12. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
13. Liu, S. et al. Leveraging Large Language Models for Generating Responses to Patient Messages. *medRxiv*, 2023.2007.2014.23292669. https://doi.org/10.1101/2023.07.14.23292669 (2023)
14. Busch, F., Hoffmann, L., Adams, L. C. & Bressem, K. K. *A systematic review of current large language model applications and biases in patient care*, https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42024504542 (2024).
15. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj* **372**, n71 (2021).
16. Ouzzani, M., Hammady, H., Fedorowicz, Z. & Elmagarmid, A. Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews* **5**, 210 (2016).
17. *Data extraction form*, https://docs.google.com/forms/d/e/1FAIpQLScFwE5KaOugxX_xXtt9Y6fbBhV4s77S9cWRdVuiHh34vmArkQ/viewform (2024).
18. Hong, Q. N. et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information* **34**, 285–291 (2018).
19. Hong, Q. N., Pluye, P., Bujold, M. & Wassef, M. Convergent and sequential synthesis designs: implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. *Syst Rev* **6**, 61 (2017).
20. Thomas, J. & Harden, A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology* **8**, 45 (2008).
21. Dedoose Version 9.2.4, cloud application for managing, analyzing, and presenting qualitative and mixed method research data (Los Angeles, CA, 2024).

22. Savage, T., Wang, J. & Shieh, L. A Large Language Model Screening Tool to Target Patients for Best Practice Alerts: Development and Validation. *JMIR Med Inform* **11**, e49886 (2023).
23. Coskun, B. N., Yagiz, B., Ocakoglu, G., Dalkilic, E. & Pehlivan, Y. Assessing the accuracy and completeness of artificial intelligence language models in providing information on methotrexate use. *Rheumatol Int*. https://doi.org/10.1007/s00296-023-05473-5 (2023)
24. Bitar, H., Babour, A., Nafa, F., Alzamzami, O. & Alismail, S. Increasing Women's Knowledge about HPV Using BERT Text Summarization: An Online Randomized Study. *Int J Environ Res Public Health* **19**. https://doi.org/10.3390/ijerph19138100 (2022)
25. Samaan, J. S. et al. Artificial Intelligence and Patient Education: Examining the Accuracy and Reproducibility of Responses to Nutrition Questions Related to Inflammatory Bowel Disease by GPT-4. *medRxiv*, 2023.2010.2028.23297723. https://doi.org/10.1101/2023.10.28.23297723 (2023)
26. Eromosele, O. B., Sobodu, T., Olayinka, O. & Ouyang, D. Racial Disparities in Knowledge of Cardiovascular Disease by a Chat-Based Artificial Intelligence Model. *medRxiv*, 2023.2009.2020.23295874. https://doi.org/10.1101/2023.09.20.23295874 (2023)
27. Johri, S. et al. Guidelines For Rigorous Evaluation of Clinical LLMs For Conversational Reasoning. *medRxiv*, 2023.2009.2012.23295399. https://doi.org/10.1101/2023.09.12.23295399 (2024)
28. Braga, A. V. N. M. et al. Use of ChatGPT in Pediatric Urology and its Relevance in Clinical Practice: Is it useful? *medRxiv*, 2023.2009.2011.23295266. https://doi.org/10.1101/2023.09.11.23295266 (2023)
29. King, R. C. et al. Appropriateness of ChatGPT in answering heart failure related questions. *medRxiv*, 2023.2007.2007.23292385. https://doi.org/10.1101/2023.07.07.23292385 (2023)
30. Huang, S. S. et al. Fact Check: Assessing the Response of ChatGPT to Alzheimer's Disease Statements with Varying Degrees of Misinformation. *medRxiv*, 2023.2009.2004.23294917. https://doi.org/10.1101/2023.09.04.23294917 (2023)
31. Hanna, J. J., Wakene, A. D., Lehmann, C. U. & Medford, R. J. Assessing Racial and Ethnic Bias in Text Generation for Healthcare-Related Tasks by ChatGPT1. *medRxiv*, 2023.2008.2028.23294730. https://doi.org/10.1101/2023.08.28.23294730 (2023)
32. Samaan, J. S. et al. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. *Arab J Gastroenterol* **24**, 145–148 (2023).
33. Patnaik, S. S. & Hoffmann, U. Quantitative evaluation of ChatGPT versus Bard responses to anaesthesia-related queries. *Br J Anaesth* **132**, 169–171 (2024).
34. Ali, H. et al. Evaluating the performance of ChatGPT in responding to questions about endoscopic procedures for patients. *iGIE* **2**, 553–559 (2023).
35. Suresh, K. et al. Utility of GPT-4 as an Informational Patient Resource in Otolaryngology. *medRxiv*, 2023.2005.2014.23289944. https://doi.org/10.1101/2023.05.14.23289944 (2023)
36. Yeo, Y. H. et al. GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis. *medRxiv*, 2023.2005.2004.23289482. https://doi.org/10.1101/2023.05.04.23289482 (2023)
37. Knebel, D. et al. Assessment of ChatGPT in the Prehospital Management of Ophthalmological Emergencies - An Analysis of 10 Fictional Case Vignettes. *Klin Monbl Augenheilkd*. https://doi.org/10.1055/a-2149-0447 (2023)
38. Zhu, L., Mou, W. & Chen, R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *Journal of Translational Medicine* **21**, 269 (2023).
39. Lahat, A., Shachar, E., Avidan, B., Glicksberg, B. & Klang, E. Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There Yet? *Diagnostics (Basel)* **13** (2023). https://doi.org/10.3390/diagnostics13111950
40. Bernstein, I. A. et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Network Open* **6**, e2330320–e2330320 (2023).
41. Rogasch, J. M. M. et al. ChatGPT: Can You Prepare My Patients for [18F]FDG PET/CT and Explain My Reports? *Journal of Nuclear Medicine*, jnumed.123.266114. https://doi.org/10.2967/jnumed.123.266114 (2023)
42. Campbell, D. J. et al. Evaluating ChatGPT Responses on Thyroid Nodules for Patient Education. *Thyroid®*. https://doi.org/10.1089/thy.2023.0491 (2023)
43. Currie, G., Robbie, S. & Tually, P. ChatGPT and patient information in nuclear medicine: GPT−3.5 versus GPT-4. *J Nucl Med Technol* **51**, 307–313 (2023).
44. Draschl, A. et al. Are ChatGPT's Free-Text Responses on Periprosthetic Joint Infections of the Hip and Knee Reliable and Useful? *J Clin Med* **12**. https://doi.org/10.3390/jcm12206655 (2023)
45. Alessandri-Bonetti, M., Liu, H. Y., Palmesano, M., Nguyen, V. T. & Egro, F. M. Online patient education in body contouring: a comparison between Google and ChatGPT. *Journal of Plastic, Reconstructive & Aesthetic Surgery* **87**, 390–402 (2023).
46. Coskun, B., Ocakoglu, G., Yetemen, M. & Kaygisiz, O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? *Urology* **180**, 35–58 (2023).
47. Durairaj, K. K. et al. Artificial Intelligence Versus Expert Plastic Surgeon: Comparative Study Shows ChatGPT "Wins" Rhinoplasty Consultations: Should We Be Worried? *Facial Plastic Surgery & Aesthetic Medicine*. https://doi.org/10.1089/fpsam.2023.0224 (2023)
48. Kianian, R., Sun, D., Crowell, E. L. & Tsui, E. The Use of Large Language Models to Generate Education Materials about Uveitis. *Ophthalmol Retina* https://doi.org/10.1016/j.oret.2023.09.008 (2023).
49. Seth, I. et al. Exploring the Role of a Large Language Model on Carpal Tunnel Syndrome Management: An Observation Study of ChatGPT. *J Hand Surg Am* **48**, 1025–1033 (2023).
50. Inojosa, H. et al. Can ChatGPT explain it? Use of artificial intelligence in multiple sclerosis communication. *Neurological Research and Practice* **5**, 48 (2023).
51. Lyons, R. J., Arepalli, S. R., Fromal, O., Choi, J. D. & Jain, N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. *Can J Ophthalmol*. https://doi.org/10.1016/j.jcjo.2023.07.016 (2023)
52. Babayiğit, O., Tastan Eroglu, Z., Ozkan Sen, D. & Ucan Yarkac, F. Potential use of ChatGPT for patient information in periodontology: a descriptive pilot study. *Cureus* **15**, e48518 (2023).
53. Mondal, H., Dash, I., Mondal, S. & Behera, J. K. ChatGPT in answering queries related to lifestyle-related diseases and disorders. *Cureus* **15**, e48296 (2023).
54. Kim, H. W., Shin, D. H., Kim, J., Lee, G. H. & Cho, J. W. Assessing the performance of ChatGPT's responses to questions related to epilepsy: a cross-sectional study on natural language processing and medical information retrieval. *Seizure* **114**, 1–8 (2023).
55. Song, H. et al. Evaluating the performance of different large language models on health consultation and patient education in urolithiasis. *J Med Syst* **47**, 125 (2023).
56. Zalzal, H. G., Abraham, A., Cheng, J. H. & Shah, R. K. Can ChatGPT help patients answer their otolaryngology questions? *Laryngoscope Investigative Otolaryngology*. https://doi.org/10.1002/lio2.1193 (2023)
57. Chervenak, J., Lieman, H., Blanco-Breindel, M. & Jindal, S. The promise and peril of using a large language model to obtain clinical

information: ChatGPT performs strongly as a fertility counseling tool with limitations. *Fertility and Sterility* **120**, 575–583 (2023).
58. Bushuven, S. et al. ChatGPT, can you help me save my child's life?" - diagnostic accuracy and supportive capabilities to lay rescuers by ChatGPT in prehospital basic life support and paediatric advanced life support cases - an in-silico analysis. *J Med Syst* **47**, 123 (2023).
59. Jeblick, K. et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European Radiology*. https://doi.org/10.1007/s00330-023-10213-1 (2023)
60. Samaan, J. S. et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obes Surg* **33**, 1790–1796 (2023).
61. Zhou, J. M., Li, T. Y., Fong, S. J., Dey, N. & Crespo, R. G. Exploring ChatGPT's potential for consultation, recommendations and report diagnosis: gastric cancer and gastroscopy reports' case. *International Journal of Interactive Multimedia and Artificial Intelligence* **8**, 7–13 (2023).
62. Oniani, D. et al. Toward improving health literacy in patient education materials with neural machine translation models. *AMIA Jt Summits Transl Sci Proc* **2023**, 418–426 (2023).
63. Hernandez, C. A. et al. The future of patient education: ai-driven guide for type 2 diabetes. *Cureus* **15**, e48919 (2023).
64. Kuşcu, O., Pamuk, A. E., Sütay Süslü, N. & Hosal, S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? *Front Oncol* **13**, 1256459 (2023).
65. Biswas, S., Logan, N. S., Davies, L. N., Sheppard, A. L. & Wolffsohn, J. S. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic Physiol Opt* **43**, 1562–1570 (2023).
66. Chiesa-Estomba, C. M. et al. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. *Eur Arch Otorhinolaryngol*. https://doi.org/10.1007/s00405-023-08104-8 (2023)
67. Decker, H. et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Netw Open* **6**, e2336997 (2023).
68. Kaarre, J. et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. *Knee Surg Sports Traumatol Arthrosc* **31**, 5190–5198 (2023).
69. Ferreira, A. L., Chu, B., Grant-Kels, J. M., Ogunleye, T. & Lipoff, J. B. Evaluation of ChatGPT dermatology responses to common patient queries. *JMIR Dermatol* **6**, e49280 (2023).
70. Truhn, D. et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. *Sci Rep* **13**, 20159 (2023).
71. Hurley, E. T. *et al*. Evaluation High-Quality of Information from ChatGPT (Artificial Intelligence-Large Language Model) Artificial Intelligence on Shoulder Stabilization Surgery. *Arthroscopy*. https://doi.org/10.1016/j.arthro.2023.07.048 (2023)
72. Cankurtaran, R. E., Polat, Y. H., Aydemir, N. G., Umay, E. & Yurekli, O. T. Reliability and usefulness of ChatGPT for inflammatory bowel diseases: an analysis for patients and healthcare professionals. *Cureus* **15**, e46736 (2023).
73. Birkun, A. A. & Gautam, A. Large language model (LLM)-powered chatbots fail to generate guideline-consistent content on resuscitation and may provide potentially harmful advice. *Prehosp Disaster Med* **38**, 757–763 (2023).
74. Pushpanathan, K. et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. *iScience* **26**, 108163 (2023).
75. Shao, C. Y. et al. Appropriateness and comprehensiveness of using ChatGPT for perioperative patient education in thoracic surgery in different language contexts: survey study. *Interact J Med Res* **12**, e46900 (2023).
76. Vaira, L. A. et al. Accuracy of ChatGPT-Generated Information on Head and Neck and Oromaxillofacial Surgery: A Multicenter Collaborative Analysis. *Otolaryngol Head Neck Surg*. https://doi.org/10.1002/ohn.489 (2023)
77. Chen, S. et al. Use of artificial intelligence Chatbots for cancer treatment information. *JAMA Oncol* **9**, 1459–1462 (2023).
78. Bellinger, J. R. et al. BPPV Information on Google Versus AI (ChatGPT). *Otolaryngol Head Neck Surg*. https://doi.org/10.1002/ohn.506 (2023)
79. Nielsen, J. P. S., von Buchwald, C. & Grønhøj, C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. *Acta Otolaryngol* **143**, 779–782 (2023).
80. Sezgin, E., Chekeni, F., Lee, J. & Keim, S. Clinical accuracy of large language models and google search responses to postpartum depression questions: cross-sectional study. *J Med Internet Res* **25**, e49240 (2023).
81. Floyd, W. et al. Current Strengths and Weaknesses of ChatGPT as a Resource for Radiation Oncology Patients and Providers. *International Journal of Radiation Oncology, Biology, Physics*. https://doi.org/10.1016/j.ijrobp.2023.10.020 (2023)
82. Uz, C. & Umay, E. "Dr ChatGPT": is it a reliable and useful source for common rheumatic diseases? *Int J Rheum Dis* **26**, 1343–1349 (2023).
83. Athavale, A., Baier, J., Ross, E. & Fukaya, E. The potential of chatbots in chronic venous disease patient management. *JVS Vasc Insights* **1**. https://doi.org/10.1016/j.jvsvi.2023.100019 (2023)
84. Li, Y. et al. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus* **15**, e40895 (2023).
85. Seth, I. et al. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: an observational study. *Aesthet Surg J Open Forum* **5**, ojad084 (2023).
86. Lockie, E. & Choi, J. Evaluation of a chat GPT generated patient information leaflet about laparoscopic cholecystectomy. *ANZ J Surg*. https://doi.org/10.1111/ans.18834 (2023)
87. Haver, H. L., Lin, C. T., Sirajuddin, A., Yi, P. H. & Jeudy, J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. *AJR Am J Roentgenol* **221**, 701–704 (2023).
88. Li, H. et al. Decoding radiology reports: potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports. *Clin Imaging* **101**, 137–141 (2023).
89. Scheschenja, M. et al. Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: a comparative analysis. *CardioVascular and Interventional Radiology* **47**, 245–250 (2024).
90. Gordon, E. B. et al. Enhancing patient communication With Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. *J Am Coll Radiol* **21**, 353–359 (2024).
91. Stroop, A. et al. Large language models: Are artificial intelligence-based chatbots a reliable source of patient information for spinal surgery? *Eur Spine J*. https://doi.org/10.1007/s00586-023-07975-z (2023)
92. Coraci, D. et al. ChatGPT in the development of medical questionnaires. The example of the low back pain. *Eur J Transl Myol* **33**. https://doi.org/10.4081/ejtm.2023.12114 (2023)
93. Ye, C., Zweck, E., Ma, Z., Smith, J. & Katz, S. Doctor Versus Artificial Intelligence: Patient and Physician Evaluation of Large Language Model Responses to Rheumatology Patient Questions in a Cross-Sectional Study. *Arthritis & Rheumatology* **n/a** https://doi.org/10.1002/art.42737
94. Mohammad-Rahimi, H. et al. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. *Int Endod J* **57**, 305–314 (2024).

95. Hermann, C. E. et al. Let's chat about cervical cancer: Assessing the accuracy of ChatGPT responses to cervical cancer questions. *Gynecologic Oncology* **179**, 164–168 (2023).
96. Kerbage, A. et al. Accuracy of ChatGPT in Common Gastrointestinal Diseases: Impact for Patients and Providers. *Clin Gastroenterol Hepatol*. https://doi.org/10.1016/j.cgh.2023.11.008 (2023)
97. Shiraishi, M. et al. Generating Informed Consent Documents Related to Blepharoplasty Using ChatGPT. *Ophthalmic Plast Reconstr Surg*. https://doi.org/10.1097/iop.0000000000002574 (2023)
98. Barclay, K. S. et al. Quality and Agreement With Scientific Consensus of ChatGPT Information Regarding Corneal Transplantation and Fuchs Dystrophy. *Cornea*. https://doi.org/10.1097/ico.0000000000003439 (2023)
99. Qarajeh, A. et al. AI-powered renal diet support: performance of ChatGPT, Bard AI, and Bing Chat. *Clin Pract* **13**, 1160–1172 (2023).
100. Chowdhury, M. et al. *Can Large Language Models Safely Address Patient Questions Following Cataract Surgery?*, (2023).
101. Singer, M. B., Fu, J. J., Chow, J. & Teng, C. C. Development and evaluation of aeyeconsult: a novel ophthalmology chatbot leveraging verified textbook knowledge and GPT-4. *J Surg Educ* **81**, 438–443 (2024).
102. Xie, Y. et al. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. *Aesthetic Plast Surg* **47**, 1985–1993 (2023).
103. Nastasi, A. J., Courtright, K. R., Halpern, S. D. & Weissman, G. E. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. *Sci Rep* **13**, 17885 (2023).
104. Biswas, M., Islam, A., Shah, Z., Zaghouani, W. & Brahim Belhaouari, S. *Can ChatGPT be Your Personal Medical Assistant?*, (2023).
105. Panagoulias, D., Palamidas, F., Virvou, M. & Tsihrintzis, G. *Evaluating the Potential of LLMs and ChatGPT on Medical Diagnosis and Treatment*. (2023).
106. Chandra, A., Davis, M. J., Hamann, D. & Hamann, C. R. Utility of allergen-specific patient-directed handouts generated by chat generative pretrained transformer. *Dermatitis* **34**, 448 (2023).
107. Hung, Y.-C., Chaker, S., Sigel, M., Saad, M. & Slater, E. Comparison of patient education materials generated by chat generative pre-trained transformer versus experts: an innovative way to increase readability of patient education materials. *Annals of Plastic Surgery* **91**, 409–412 (2023).
108. Capelleras, M., Soto-Galindo, G. A., Cruellas, M. & Apaydin, F. ChatGPT and Rhinoplasty Recovery: An Exploration of AI's Role in Postoperative Guidance. *Facial Plast Surg* (2024). https://doi.org/10.1055/a-2219-4901
109. Scquizzato, T. et al. Testing ChatGPT ability to answer laypeople questions about cardiac arrest and cardiopulmonary resuscitation. *Resuscitation* **194**, 110077 (2024).
110. Kuckelman, I. J. et al. Assessing AI-powered patient education: a case study in radiology. *Acad Radiol* **31**, 338–342 (2024).
111. Sulejmani, P. et al. A large language model artificial intelligence for patient queries in atopic dermatitis. *J Eur Acad Dermatol Venereol*. https://doi.org/10.1111/jdv.19737 (2024)
112. Currie, G. & Barry, K. ChatGPT in nuclear medicine education. *Journal of Nuclear Medicine Technology* **51**, 247–254 (2023).
113. Currie, G. M. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? *Seminars in Nuclear Medicine* **53**, 719–730 (2023).
114. Li, J., Dada, A., Puladi, B., Kleesiek, J. & Egger, J. ChatGPT in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine* **245**, 108013 (2024).
115. Liu, F. W. & Hu, C. Exploring Vulnerabilities and Protections in Large Language Models: A Survey. *arXiv preprint arXiv:2406.00240* (2024).
116. Jin, M. et al. Health-LLM: Personalized Retrieval-Augmented Disease Prediction Model. *arXiv preprint arXiv:2402.00746* (2024).
117. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).
118. Brin, D. et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Scientific Reports* **13**, 16492 (2023).
119. Jung, L. B. et al. ChatGPT passes German State examination in medicine with picture questions omitted. *Dtsch Arztebl Int* **120**, 373–374 (2023).
120. Bhayana, R., Krishna, S. & Bleakney, R. R. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology* **307**, e230582 (2023).
121. Singhal, K. et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).
122. Dorfner, F. J. et al. Biomedical Large Languages Models Seem not to be Superior to Generalist Models on Unseen Medical Data. *arXiv preprint arXiv:2408.13833* (2024).
123. Kung, T. H. et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health* **2**, e0000198 (2023).
124. Kapoor, S., Henderson, P. & Narayanan, A. Promises and pitfalls of artificial intelligence for legal applications. *arXiv preprint arXiv:2402.01656* (2024).
125. Ouyang, S., Zhang, J. M., Harman, M. & Wang, M. LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation. *arXiv preprint arXiv:2308.02828* (2023).
126. Atil, B. et al. LLM Stability: A detailed analysis with some surprises. *arXiv preprint arXiv:2408.04667* (2024).
127. Weis, B. Health Literacy: A Manual for Clinicians. Chicago, IL: American Medical Association, American Medical Foundation; 2003. National Institutes of Health. How to Write Easy to Read Health Materials: National Library of Medicine Website. *How to Write Easy to Read Health Materials: National Library of Medicine Website*
128. Amann, J. et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making* **20**, 310 (2020).
129. Gerke, S., Minssen, T. & Cohen, G. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare*, 295-336. https://doi.org/10.1016/b978-0-12-818438-7.00012-5 (2020)
130. Tobias, J. S. & Souhami, R. L. Fully informed consent can be needlessly cruel. *Bmj* **307**, 1199–1201 (1993).
131. Pietrzykowski, T. & Smilowska, K. The reality of informed consent: empirical studies on patient comprehension—systematic review. *Trials* **22**, 57 (2021).
132. Kesselheim, A. S., Connolly, J., Rogers, J. & Avorn, J. Mandatory disclaimers on dietary supplements do not reliably communicate the intended issues. *Health Affairs* **34**, 438–446 (2015).
133. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* **25**, 44–56 (2019).
134. Raj, H., Gupta, V., Rosati, D. & Majumdar, S. Semantic consistency for assuring reliability of large language models. *arXiv preprint arXiv:2308.09138* (2023).
135. Zhou, Y. et al. Large Language Models Are Human-Level Prompt Engineers. *ArXiv* **abs/2211.01910** (2022).
136. Zhang, Z. et al. Certified Robustness for Large Language Models with Self-Denoising. *ArXiv* **abs/2307.07171** (2023).
137. Ullah, E., Parwani, A., Baig, M. M. & Singh, R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. *Diagn Pathol* **19**, 43 (2024).

138. Navigli, R., Conia, S. & Ross, B. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality* **15**, 1–21 (2023).
139. Deng, G. et al. Jailbreaker: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715* (2023).
140. Acerbi, A. & Stubbersfield, J. M. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences* **120**, e2313790120 (2023).
141. Yang, Y., Liu, Y. & Naghizadeh, P. Adaptive data debiasing through bounded exploration. *Advances in Neural Information Processing Systems* **35**, 1516–1528 (2022).
142. Gallegos, I. O. et al. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 1-83. https://doi.org/10.1162/coli_a_00524 (2024)
143. Grari, V., Laugel, T., Hashimoto, T., Lamprier, S. & Detyniecki, M. On the Fairness ROAD: Robust Optimization for Adversarial Debiasing. *arXiv preprint arXiv:2310.18413* (2023).
144. Hofmann, V., Kalluri, P. R., Jurafsky, D. & King, S. AI generates covertly racist decisions about people based on their dialect. *Nature* **633**, 147–154 (2024).
145. Ayoub, N. F., Lee, Y. J., Grimm, D. & Divi, V. Head-to-Head Comparison of ChatGPT Versus Google Search for Medical Knowledge Acquisition. *Otolaryngol Head Neck Surg*. https://doi.org/10.1002/ohn.465 (2023)
146. Yao, Y. et al. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. *High-Confidence Computing* **4**, 100211 (2024).
147. Tierney, A. A. et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst* **5**, CAT.23.0404 (2024).
148. Formosa, P., Rogers, W., Griep, Y., Bankins, S. & Richards, D. Medical AI and human dignity: contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts. *Computers in Human Behavior* **133**, 107296 (2022).
149. Moy, S. et al. Patient perspectives on the use of artificial intelligence in health care: a scoping review. *J Patient Cent Res Rev* **11**, 51–62 (2024).
150. Union, C. O. T. E. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - Analysis of the final compromise text with a view to agreement, (2024).
151. Busch, F. et al. Navigating the european union artificial intelligence act for healthcare. *npj Digital Medicine* **7**, 210 (2024).

Acknowledgements

This research is funded by the European Union (101079894). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them. The funding had no role in the study design, data collection and analysis, manuscript preparation, or decision to publish.

Author contributions

Conceptualization: FB, LH, CR, EHCvD, RK, EOP, MRM, LS, MH, JNK, DT, RC, LCA, KKB; Project administration: FB; Resources: FB, LCA, KKB; Software: FB, LCA, KKB; Data curation: FB, LH, CR; Formal analysis: FB, LH, CR, LCA, KKB; Investigation: FB, LH, CR, LCA, KKB; Methodology: FB; Supervision: FB, LCA, KKB; Validation: FB, LH, CR, EHCvD, RK, EOP, MRM, LS, MH, JNK, DT, RC, LCA, KKB; Visualization: FB, LCA; Writing – original draft preparation: FB, LH, LCA, KKB; Writing – review & editing: FB, LH, CR, EHCvD, RK, EOP, MRM, LS, MH, JNK, DT, RC, LCA, KKB.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

JNK declares consulting services for Owkin, France; DoMore Diagnostics, Norway; Panakeia, UK, and Scailyte, Basel, Switzerland; furthermore JNK holds shares in Kather Consulting, Dresden, Germany; and StratifAI GmbH, Dresden, Germany, and has received honoraria for lectures and advisory board participation by AstraZeneca, Bayer, Eisai, MSD, BMS, Roche, Pfizer and Fresenius. DT holds shares in StratifAI GmbH, Dresden, Germany and has received honoraria for lectures by Bayer. KKB reports grants from the European Union (101079894) and Wilhelm-Sander Foundation; participation on a Data Safety Monitoring Board or Advisory Board for the EU Horizon 2020 LifeChamps project (875329) and the EU IHI Project IMAGIO (101112053); speaker Fees for Canon Medical Systems Corporation and GE HealthCare. RK receives medical consultancy fees from Odin Vision. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-024-00717-2.

Correspondence and requests for materials should be addressed to Felix Busch.

Peer review information *Communications Medicine* thanks Dmitry Scherbakov and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

[1]School of Medicine and Health, Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, TUM University Hospital, Technical University of Munich, Munich, Germany. [2]Department of Neuroradiology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany. [3]Department of Ophthalmology, Leiden University Medical Center, Leiden, The Netherlands. [4]Department of Ophthalmology, Sir Charles Gairdner Hospital, Perth, Australia. [5]Division of Surgery and Interventional Sciences, University College London, London, United Kingdom. [6]One Health Research Group,

Faculty of Health Science, Universidad de Las Américas, Quito, Ecuador. [7]Department of Radiology, Azienda Ospedaliero Universitaria (A.O.U.), Cagliari, Italy. [8]School of Medicine and Health, Institute for Cardiovascular Radiology and Nuclear Medicine, German Heart Center Munich, TUM University Hospital, Technical University of Munich, Munich, Germany. [9]Department of Medical Oncology, National Center for Tumor Diseases (NCT), Heidelberg University Hospital, Heidelberg, Germany. [10]Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany. [11]Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany. [12]Department of Medicine, Surgery and Dentistry, University of Salerno, Baronissi, Italy. [13]These authors contributed equally: Lisa C. Adams, Keno K. Bressem. ✉e-mail: felix.busch@tum.de