Systematic Review of Large Language Models for Patient Care: Current Applications and Challenges

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

[1] Department of Neuroradiology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany

[2] Department of Ophthalmology, Leiden University Medical Center, Leiden, The Netherlands

[3] Department of Ophthalmology, Sir Charles Gairdner Hospital, Perth, Australia

[4] Division of Surgery and Interventional Sciences, University College London, London, United Kingdom

[5] One Health Research Group, Faculty of Health Science, Universidad de Las Américas, Quito, Ecuador

[6] Department of Radiology, Technical University of Munich, Munich, Germany

[7] Department of Radiology, Azienda Ospedaliero Universitaria (A.O.U.), Cagliari, Italy

[8] Institute for Radiology and Nuclear Medicine, German Heart Center Munich, 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

[*] Correspondence to: Felix Busch, MD; Address: Department of Neuroradiology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Charitépl. 1, 10117 Berlin, Germany; E-mail: felix.busch@charite.de

[#] These authors contributed equally to this work.

Abstract

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 using a data-driven convergent synthesis approach. We searched 5 databases for qualitative, quantitative, and mixed methods articles on LLMs in patient care published between 2022 and 2023. From 4,349 initial records, 89 studies across 29 medical specialties were included, primarily examining models based on the GPT-3.5 (53.2%, n=66 of 124 different LLMs examined per study) and GPT-4 (26.6%, n=33/124) architectures in medical question answering, 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 included 6 second-order and 12 third-order codes, such as lack of medical domain optimization, data transparency, and accessibility issues, while output limitations included 9 second-order and 32 third-order codes, for example, non-reproducibility, non-comprehensiveness, incorrectness, unsafety, and bias. In conclusion, this study is the first review to systematically map LLM applications and limitations in patient care, providing a foundational framework and taxonomy for their implementation and evaluation in healthcare settings.

Keywords: Artificial Intelligence; Bias; Health Personnel; Medical Informatics; Natural Language Processing; Patient Care

1. Introduction

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 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]

In contrast to applications primarily aimed at healthcare professionals, LLMs could also be used to promote patient 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.

However, despite the growing body of research and the clear potential of LLMs, there is a gap in terms of systematized information towards their use in patient care. To date, 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.

Therefore, 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.

2. 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-Analyses (PRISMA) guidelines.[14,15]

2.1 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.

2.2 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 Supplementary Section 1). 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.

2.3 Data analysis

Due to the diversity of investigated outcomes and study designs we sought to include, including qualitative, quantitative, and mixed methods, 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.

3. Results

3.1 Screening results

Of the 4,349 reports identified, 2,991 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 Section 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 Figure 1.

3.2 Characteristics of included studies

Table 1 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 Table 2). 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%), 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]

3.3 Applications of Large Language Models

An overview of the presence of codes for each study is provided in Supplementary Section 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=19/89, 21.4%)[24,31,43,48,49,57,59,62,67,70,79,88-91,97,102,106,107], 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], gastro-enterology (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 Figure 2.

3.4 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.

3.4.1 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, 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 dialogue capabilities (n=10/89)[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].

3.4.2 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].

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 included nine concepts related to generic/non-personalized output (n=34/89, 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]. 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). However, numerous 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 beliefs 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).

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). Finally, 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]. Figure 3 illustrates the hierarchical tree structure and quantity of the codes derived from the thematic synthesis of limitations.

4. 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 their future implementation and evaluation 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.

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, the use of proprietary models such as ChatGPT in the biomedical field raises concerns because the limited access to the underlying algorithms, training data, and data processing and storage mechanisms makes them untransparent and, thus, significantly limits their applicability in healthcare.[114] Furthermore, the integration of proprietary models into patient care applications makes one susceptible to performance changes associated with model updates, which may break existing functionalities and lead to

harmful outcomes for patients. Therefore, especially in the biomedical field, open-source models such as BioMistral may offer a viable solution.[6] 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.

We identified several key limitations regarding the design and output. Not surprisingly, many reports noted the limitation that the LLMs studied were not optimized for the medical domain. One possible solution to this limitation may be to provide medical knowledge during inference using RAG.[115] 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.[116-119] 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.[120] 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.

It should also be noted that passing these exams does not equate to the practical competence required of a healthcare provider.[121] 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.[122] In fact, these types of tests can be misleading in estimating the model's true abilities in terms of comprehension or analytical skills.

Many studies have reported limitations in the output related to comprehensiveness, safety, correctness, reproducibility, and dependence of the output on the input/prompt and environment. Specifically, for correctness, 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] On the other hand, a minority of studies have identified biases, for example, reflecting the unequal representation of certain content or the biases inherent in human-generated text in the training data.[123] This may indicate that the implemented safeguards are effective. However, not much is known about the technology and developer policies of proprietary LLMs, and previous work has shown that automated jailbreak generation is possible across various commercial LLM chatbots.[124] This also mirrors our concept of data-related limitations, particularly regarding the handling of sensitive health information. Together with the limited transparency about the origin of the training data and the unexplainable and non-deterministic nature of the output, this raises 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.[125]

In addition, low health literacy due to the identified limitations in comprehensiveness, including the generation of content with high complexity and an inappropriate reading level, which was above the 6th-grade level recommended by the American Medical Association (AMA) in almost all studies analyzed, may further limit their utility for patient information.[126] Overall, this can lead to results that are misleading and harmful, as described in many of the reports in our review. In addition to advances in the development of LLMs and the focus on open source, it will therefore be necessary to develop and implement a well-validated scale to determine the quality and safety of LLM outputs in medical practice, such as the recent effort made to adopt the widely recognized Physician Documentation Quality Instrument (PDQI-9) for the assessment of AI transcripts and clinical summaries.[127]

Finally, the implementation of regulatory mandates like the forthcoming European Union AI Act and the associated challenges faced by generative AI and LLMs, 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. 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.[128]

4.1 Limitations

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, 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. Third, 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. Continued updates will be essential to monitor emerging areas and limitations in this rapidly evolving field.

5. Conclusion

In conclusion, this review provides a systematic overview of current LLM applications and limitations in patient care. Our conceptual synthesis provides a structured taxonomy that may lay the groundwork for both the implementation and critical evaluation of LLMs in healthcare settings.

6. Declarations

6.1 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.

6.2 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.

6.3 Author contributions

Conceptualization: FB, 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.

6.4 Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

7. Figures

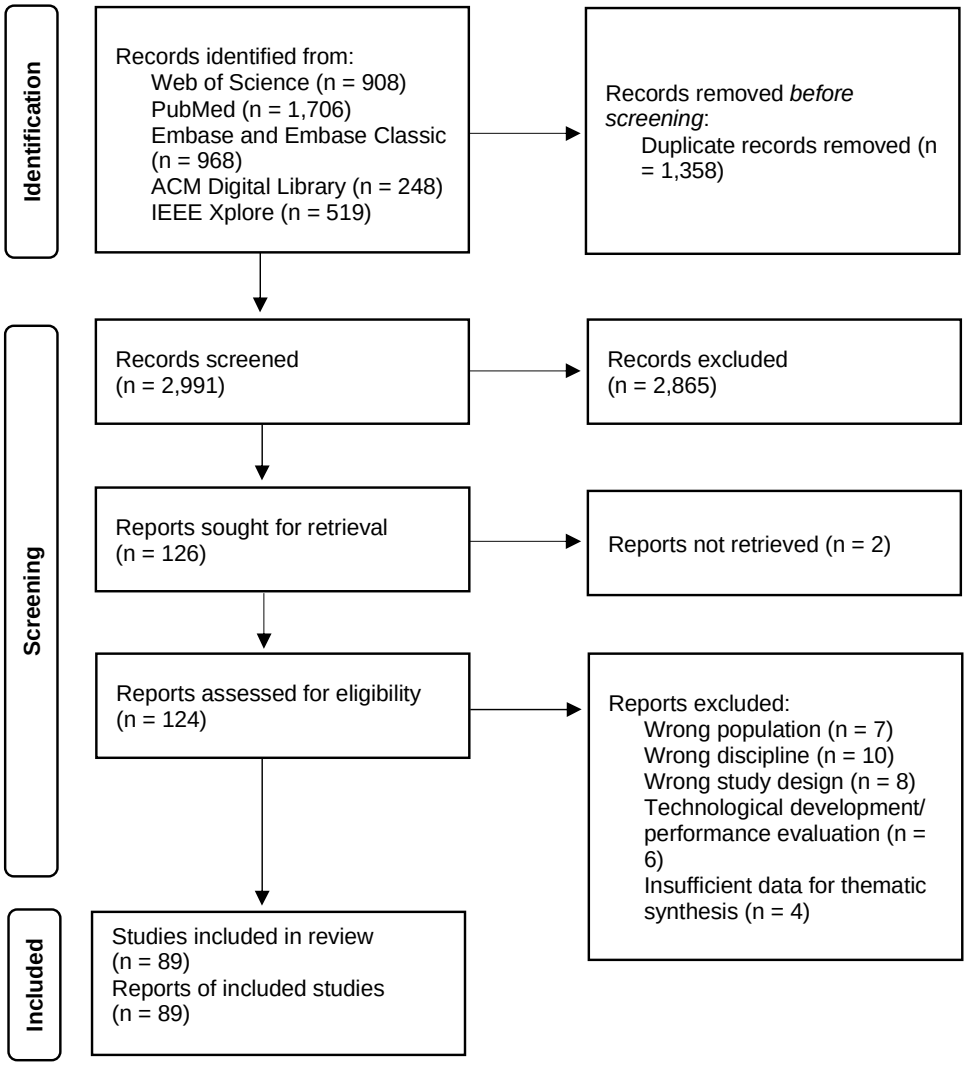


Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram.

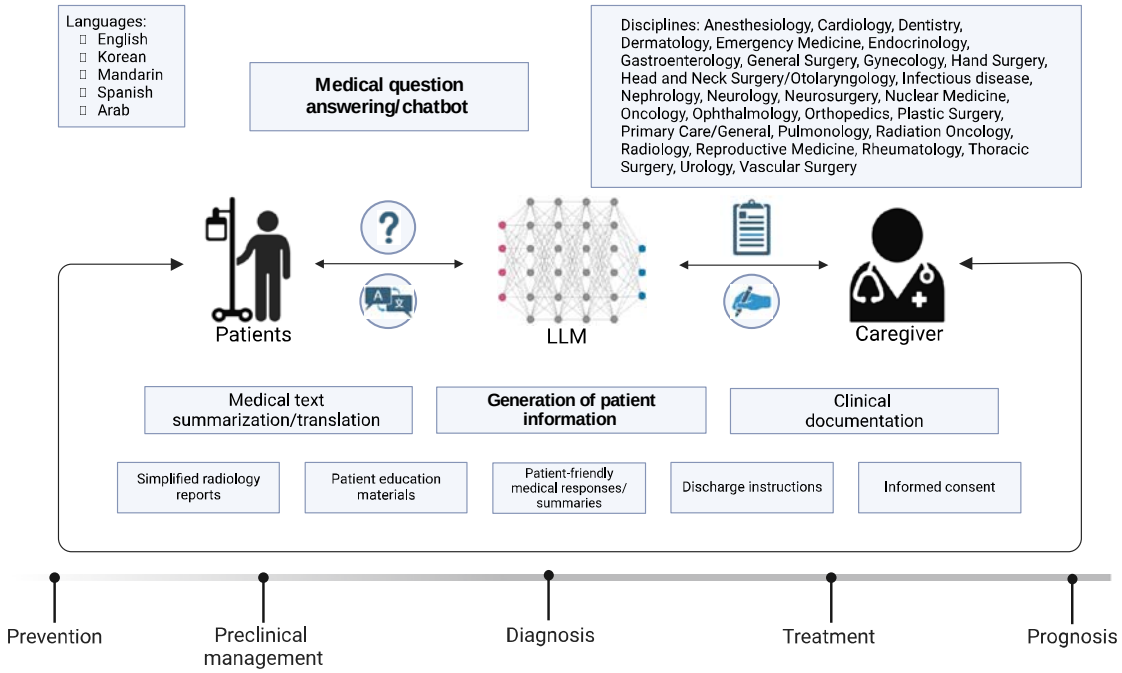


Figure 2. Schematic illustration of the identified concepts for the application of large language models (LLMs) in patient care.

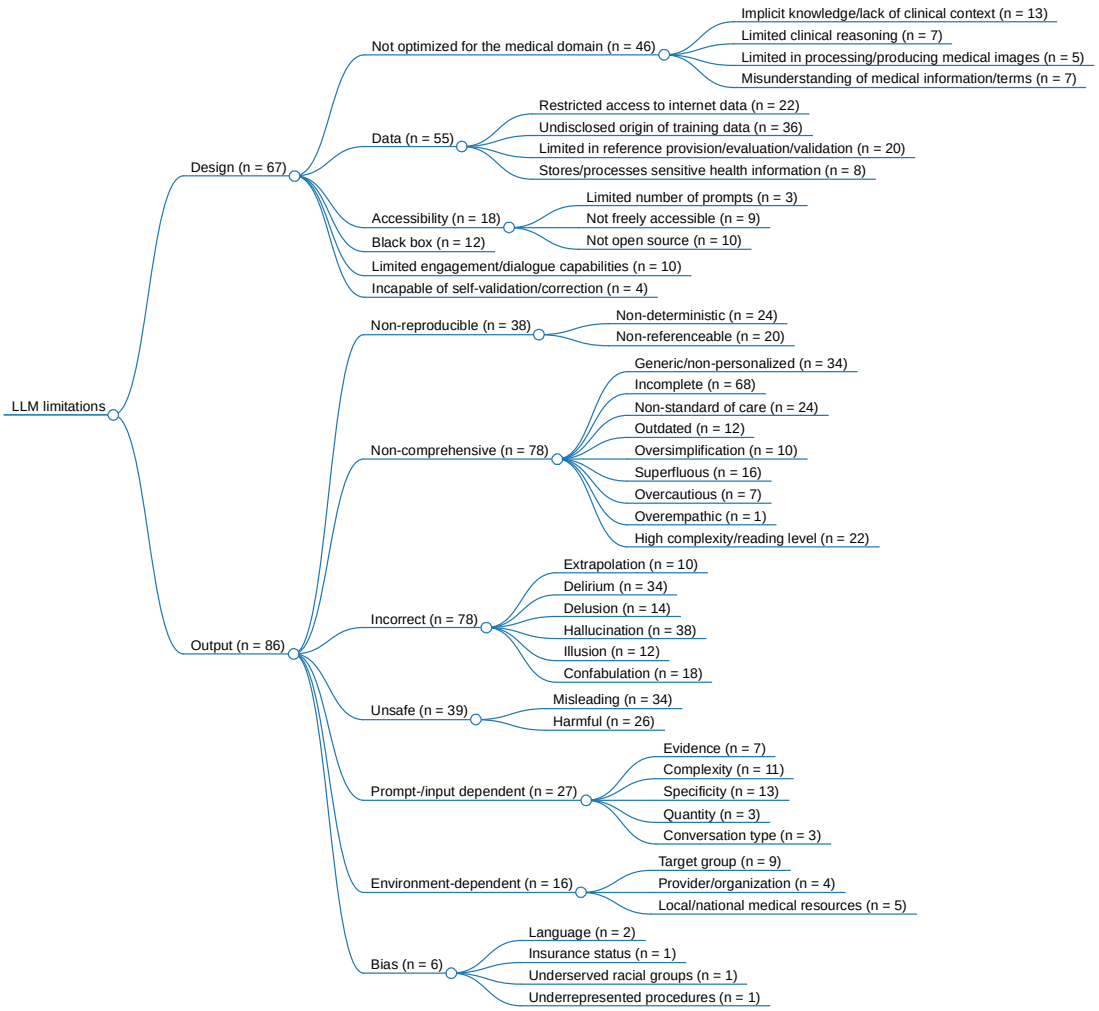


Figure 3. Illustration of the hierarchical tree structure for the thematic synthesis of large language model (LLM) limitations in patient care, including the presence of codes for each concept.

8. Tables

Table 1. Overview of included studies and corresponding authors, year of publication, affiliation countries of authors, study design, medical specialty, purpose of study, large language model (LLM)/tool examined, target user, evaluation/setting, main outcome, and conclusion.

Authors, year; country	Study design	Specialty; purpose	LLM/tool; target user	Evaluation/setting	Main outcome	Overall conclusion
1. Samaan et al., 2023; USA, UK[25]	Quantitative	Gastroenterology; GPT-4's accuracy and reproducibility in responding to patient queries on inflammatory bowel disease (IBD) nutrition	GPT-4; patients	88 questions from adult patients in a tertiary hospital, Facebook, Reddit; response assessment by 2 IBD dietitians	Accuracy: 83%, comprehensiveness: 69%, mixed accuracy: 17%, completely incorrect: 0%	GPT-4 is a promising tool for IBD patients seeking nutrition-related information
2. Eromosele et al., 2023; USA[26]	Quantitative	Cardiology; GPT-3.5 knowledge of cardiovascular disease (CVD) racial disparities	GPT-3.5; patients; caregivers	60 questions were prompted 3 times; response assessment by a team of cardiologists and other specialized clinicians	Appropriate knowledge: 63.4%, inappropriate: 33.3%, unreliable: 3.3%; 91%/79% of incorrect/hedging responses involved CVD disparities affecting a minority or underserved racial group	GPT-3.5 has satisfactory but suboptimal knowledge of CVD racial disparities
3. Johri et al., 2023; USA[27]	Quantitative	Dermatology; GPT-4's and GPT-3.5's clinical reasoning capabilities using a multi-agent conversational framework	GPT-3.5/GPT-4; patients/caregivers	140 skin cancer cases, 100 from an online dermatology website, 40 new cases by 3 dermatology residents; diagnostic accuracy of GPT-4 and GPT-3.5 in conversational versus static settings	91.9%/83.3% accuracy of GPT-4/GPT-3.5 for static vignettes, 85.4%/72.4% for cases in conversational format; GPT-3.5 accuracy improves through conversational summarization (72.4% to 81%); increasing multiple choice (MC)-options lead to decreasing accuracy (GPT-4: 85.4% to 57.2%; GPT-3.5: 72.4% to 20.1%)	GPT-4/GPT-3.5 have limitations in integrating details from conversational interactions
4. Braga et al., 2023; Brazil, Canada, USA[28]	Qualitative	Pediatric Urology; GPT-3.5's reliability in concept description and usefulness for decision-making in urology	GPT-3.5; patients/caregivers	3 queries each for primary megaureter, enuresis, and vesicoureteral reflux were prompted twice; qualitative evaluation by 3 specialists	GPT-3.5's responses partly contain accurate and pertinent information, but most are insufficient and misleading; better performance in less complex questions	GPT-3.5 lacks clinical experience and judgment, potentially providing false information
5. King et al., 2023; USA[29]	Quantitative	Cardiology; GPT-4's and GPT-3.5's accuracy and reproducibility in answering heart failure questions; performance between GPT-3.5 and GPT-4	GPT-3.5/GPT-4; patients	107 questions related to heart failure from medical societies/institutions and Facebook support groups were prompted twice; assessment by 2 board-certified cardiologists	GPT-4 had highest accuracy in basic knowledge/management: 89.8%/82.9%; GPT-3.5 had highest accuracy in management/other (forecasting, procedures, support): 78.1%/94.1%; partially/incorrect answers: GPT-3.5: 1.9%/GPT-4: 0%; reproducibility: GPT-3.5: 94%/GPT-4: 100%	GPT-3.5/GPT-4 have the potential to serve as accurate and reliable resources for patients with heart failure
6. Huang et al., 2023; USA[30]	Quantitative	Neurology; GPT-4's effectiveness in identifying and explaining misinformation about Alzheimer's disease (AD) and generating audience-specific, readable explanations	GPT-4; patients/caregivers	20 myths about AD validated by 200 Amazon Mechanical Turk participants; response assessment by 11 practitioners/clinicians working primarily in geriatrics	GPT-4 identified 100% of the myths as false; readability: 81% strongly agree/agree; overall results of information retention: 82% strongly agree/agree; potential in clarifying AD misinformation: 82% strongly agree/agree	GPT-4 has potential value in mitigating AD misinformation; need for more refined/detailed explanations of disease mechanisms and treatments
7. Hanna et al., 2023; USA[31]	Quantitative	Infectiology; GPT-3.5's potential racial and ethnic bias in writing discharge instructions for HIV patients	GPT-3.5; patients/caregivers	Discharge instructions based on 100 deidentified HIV patient encounters prompted 4 times each while switching race and ethnicity; statistical assessment of polarity, subjectivity, named entity recognition (NER), Flesch Reading Ease score (readability score), Flesch-Kincaid Grade Level (readability grade), word frequency	No significant differences in generated text regarding polarity, subjectivity, NER, readability, and text length across different races/ethnicities and insurance types; statistically significant differences in word frequency across races/ethnicities and subjectivity across insurance types (commercial insurance eliciting most subjective responses)	GPT-3.5 is relatively invariant to race/ethnicity and insurance type in terms of linguistic and readability measures

8. Liu et al., 2023; USA[13]	Quantitative	Primary Care; to develop a fine-tuned LLM on messages and healthcare provider responses from a patient portal; to assess and compare responses to actual provider responses and responses from GPT-3.5/GPT-4	GPT-3.5/GPT-4, CLAIR-Short/-Long based on LLaMA-65B; patients/caregivers	10 representative questions based on a patient message framework were chosen by a primary care physician; responses by the LLMs were randomized and evaluated by 4 primary care physicians, and BERTScore values	Responses generated by GPT-3.5/4 received the highest ratings in terms of empathy, responsiveness, accuracy, and usefulness; GPT-3.5/4 and CLAIR-Long outperform CLAIR-Short and doctors' responses significantly	LLMs have great potential in generating responses to patient messages, facilitating communication between patients and primary care providers
9. Samaan et al., 2023; USA[32]	Quantitative	Gastroenterology; GPT-3.5's ability to understand and respond to liver cirrhosis patient questions in Arabic compared to English	GPT-3.5; patients	91 liver cirrhosis questions from professional societies, institutions, and Facebook support groups were translated into Arabic by 2 bilingual physicians; response grading by a bilingual board-certified transplant hepatologist	Arabic: 72.5% of questions correct, thereof 24.2% graded as comprehensive; English: 79.1% correct, 47.3% graded as comprehensive; 9.9% of the Arabic answers were rated as more accurate, 57.1% as similarly accurate, and 33% as less accurate compared to English	GPT-3.5 has the potential to serve as an additional source of information for Arabic-speaking patients with liver cirrhosis, although its performance in Arabic is less accurate than in English
10. Patnaik et al., 2023; USA[33]	Quantitative	Anesthesiology; GPT-3.5's and Bard's ability to answer anesthesia-related patient questions before surgery	GPT-3.5, Bard; patients/caregivers	11 anesthesia-related questions were collected during pre-anesthesia consultation; response evaluation using hallucination counts, readability using Flesch-Kincaid Grade Level (FKG), lexical diversity measures (MTLD), computational sentiment analysis, and Levenshtein distances	GPT-3.5 displayed no hallucination errors, Bard had a 30.3% error rate; FKG scores and MTLD were higher for GPT-3.5 compared to Bard, subjectivity scores were similar	ChatGPT was technical, precise, and descriptive, whereas Bard was conversational, adequate, and exhibited hallucinations.
11. Ali et al., 2023; USA[34]	Quantitative	Gastroenterology; GPT-3.5's and GPT-4's ability to answer questions about EGD, colonoscopy, EUS, and ERCP and provide emotional support	GPT-3.5/GPT-4; patients	113 questions on endoscopic procedures collected from professional societies or institutional websites; response grading by 2 board-certified gastroenterologists or 2 advanced endoscopists; evaluation of emotional support questions by a certified psychiatrist	Comprehensiveness: EGD: 57.9%, colonoscopy: 47.6%, EUS: 48.1%, ERCP: 44.4%; medical accuracy: highest for EGD (52.6% fully accurate), lowest for EUS (40.7% fully accurate); superfluous content: responses were predominantly concise for EGD and colonoscopy, with ERCP and EUS showing increased extraneous content; reproducibility scores: varied across domains, from 50.34% (for EUS) to 68.6% (for EGD); emotional support: GPT-4 outperformed GPT-3.5	GPT3.5/4 holds promise as a supplementary patient resource for common endoscopic procedures.
12. Suresh et al., 2023; USA[35]	Quantitative	Otolaryngology; GPT-4's utility as an informational resource for otolaryngology patients	GPT-4; patients	18 otolaryngology questions were designed based on the American Academy of Otolaryngology's Clinical Practice Guidelines; response evaluation by clinicians with expertise or subspecialty training in otolaryngology	Safe responses: 100%; Accurate responses: 78%; Comprehensive responses: 83%	GPT-4's otolaryngology advice is safe but lacks accuracy and comprehensiveness, limiting its utility as an informational resource for patients
13. Yeo et al., 2023; USA[36]	Quantitative	Gastroenterology; GPT-3.5 versus GPT-4 in understanding and responding to cirrhosis-related questions in English, Korean, Mandarin, and Spanish	GPT-3.5/GPT-4; patients	36 English liver cirrhosis questions collected from healthcare organizations and patient support groups were translated into Korean, Mandarin, and Spanish; accuracy grading based on the American Association for the Study of Liver Diseases guidelines; similarity of non-English responses graded by native-speaking hepatologists	GPT-4 showed higher response accuracy compared to GPT-3.5 across all languages; GPT-4 performed significantly better in Mandarin and Korean than GPT-3.5	GPT-4 outperformed GPT-3.5 in responding to English and non-English questions related to cirrhosis
14. Knebel et al., 2023; Germany[37]	Quantitative	Ophthalmology; GPT-3.5's performance in triaging ophthalmological emergencies	GPT-3.5; patients	10 case vignettes, derived from guideline-based prevention topics, each prompted 5 times; responses were assessed for triage accuracy, appropriateness of recommended preclinical measures, and potential harm	Triage accuracy: 93.6%; treatment accuracy: 100%; diagnosis accuracy: 61.5%; appropriate prehospital measures: 66%; potentially harmful to users/patients: 32%	GPT-3.5 should not be used as the sole primary source of information about acute ophthalmologic symptoms
15. Zhu et al., 2023; China[38]	Quantitative	Urology; LLM's utility as consultants for prostate cancer patients	GPT-3.5 free/plus, YouChat, NeevaAI, Perplexity (concise and detailed model), Chatsonic; patients	22 prostate cancer questions based on patient education guidelines from the Centers for Disease Control and Prevention and UpToDate; response evaluation based on accuracy, comprehensiveness, patient readability, humanistic care, and stability by 3 urologists	GPT-3.5 free responses had the highest accuracy (100% correct), were most comprehensive (95.5% very comprehensive), and most consistent (100%); readability was highest for GPT-3.5 free/plus (100%) compared to other LLMs	LLMs have the potential to be applied in the education and consultation of prostate cancer patients but are not yet capable of completely replacing doctors

16. Lahat et al., 2023; Israel[39]	Quantitative	Gastroenterology; GPT-3.5's performance in answering patient gastroenterological health questions	GPT-3.5; patients	110 gastroenterology questions were collected from open internet sites providing medical information to diverse patients' questions; response assessment for accuracy, clarity, up-to-date knowledge, and effectiveness by 3 gastroenterologists	Mean scores (1 to 5) for treatment questions: accuracy: 3.9, clarity: 3.9, efficacy: 3.3; mean scores for symptom questions: accuracy: 3.4, clarity: 3.7, efficacy: 3.2; mean scores for diagnostic test questions: accuracy: 3.7, clarity: 3.7, efficacy: 3.5	GPT-3.5 has potential as an information source in the field of gastroenterology, but further development is needed
17. Bernstein et al., 2023; USA[40]	Quantitative	Ophthalmology; GPT-3.5's quality of ophthalmology advice compared to ophthalmologist-written advice	GPT-3.5; patients/caregivers	200 question-answer pairs from the Eye Care Forum, with responses from American Academy of Ophthalmology-affiliated physicians; generated and original questions randomly presented to 8 board-certified ophthalmologists, assessment for incorrect information, harm likelihood and severity, medical consensus alignment	Mean accuracy of the expert panel for distinguishing between AI and human answers was 61.3% (individual rater accuracy range: 45% to 74%); no significant differences in responses containing incorrect or inappropriate information and likelihood or extent of harm between GPT-3.5 and human answers	GPT-3.5 provides ophthalmologic advice of comparable quality to that of ophthalmologists
18. Rogasch et al., 2023; Germany[41]	Quantitative	Nuclear Medicine; GPT-4's ability to answer patient questions related to [^{18}F]FDG PET/CT imaging for Hodgkin lymphoma or lung cancer	GPT-4; patients	25 tasks, including responding to 13 frequently asked patient questions/6 follow-up questions and explaining six fictitious PET/CT reports prompted 3 times; rating by 3 nuclear medicine physicians for appropriateness, helpfulness, inconsistency, validity of references	Appropriate responses: 92%; helpful: 96%; considerable inconsistencies: 16%; references fully valid: 21%	GPT-4 has the potential to offer adequate informational counseling to patients undergoing [^{18}F]FDG PET/CT imaging
19. Campbell et al., 2023; USA[42]	Quantitative	Otolaryngology; GPT-3.5's utility as an educational resource for patients on thyroid nodules	GPT-3.5; patients	30 questions on thyroid nodules were prompted 4 times using different prompting strategies: no prompting, patient-friendly prompting, 8th-grade-level prompting, and prompting for references; response grading for medical accuracy and clinical appropriateness by 2 otolaryngology resident physicians	69.2% of responses were "at least correct" and did not differ by prompting strategy; 87.5% of medical literature references were legitimate citations thereof 17.1% with incorrectly or completed falsified findings; 12.5% of references were unfindable or incorrect	GPT-3.5 answers most questions about thyroid nodules appropriately, regardless of prompting
20. Currie et al., 2023; Australia[43]	Quantitative	Nuclear Medicine; GPT-3.5's and GPT-4's ability to create patient information sheets for nuclear medicine procedures	GPT-3.5/GPT-4; patients	7 patient information sheets suitable for gaining informed consent for 7 common procedures in nuclear medicine; assessment by 3 nuclear medicine technologists or scientists for accuracy, appropriateness, currency, and fitness for the purpose	GPT-4 outperformed GPT-3.5 in accuracy, appropriateness, currency, and fitness-for-purpose but was often below the minimum standard; GPT-3.5's responses were below average for all except bone scan (which was average), GPT-4 produced higher-quality patient information sheets, with 3 classified as fit for the purpose	GPT-3.5 is ineffective for nuclear medicine patient information; GPT-4 provides more accurate patient information and may be used for informed consent
21. Draschl et al., 2023; Austria[44]	Quantitative	Orthopedics; GPT-3.5's performance in answering questions about periprosthetic joint infections of the hip and knee	GPT-3.5; patients/caregivers	27 questions from the 2018 International Consensus Meeting on Musculoskeletal Infection; response evaluation by 3 orthopedic surgeons for completeness, misleading information, errors, up-to-dateness, patient and surgeon suitability	Median completeness, up-to-dateness, patient/surgeon suitability of responses (on a 5-point Likert scale, with 5 indicating strongly agree): 4; median Likert-Scale scores for misleading or erroneous responses (with 5 indicating strongly disagree): 4	GPT-3.5 is a predominantly reliable and useful tool for orthopedic surgeons and patients in complex orthopedic questions
22. Alessandri-Bonetti et al., 2023; USA[45]	Quantitative	Plastic Surgery; GPT-3.5's potential as a viable source for patient education on body contouring compared to Google search	GPT-3.5; patients	15 questions and responses/references from the "People also ask" section of a Google search for "body contouring surgery"; 4 blinded plastic surgeons rated the answer quality of Google and GPT-3.5 using the Global Quality Score	Google responses were rated as poor quality with limited usefulness to patients (mean Likert score: 2.55); GPT-3.5 responses were rated as higher quality and more useful to patients (mean Likert score: 4.38); 33% of GPT-3.5 responses did not provide references when asked; 6% of references were inaccessible or linked to unrelated sites	GPT-3.5 outperformed Google search and can be a useful tool for patient education on body contouring
23. Capelleras et al., 2024; Turkey, Spain[108]	Qualitative	Plastic Surgery; GPT-3.5's potential in providing postoperative guidance during rhinoplasty recovery	GPT-3.5; patients	8 standardized questions were formulated based on the Rhinobase 2.0 database; qualitatively response assessment for recurring themes, patterns, and trends related to rhinoplasty recovery	GPT-3.5's responses guide common concerns after rhinoplasty, including swelling, emotional adjustment, asymmetry, breathing difficulties, pain, skin color changes, bleeding, and numbness; GPT-3.5 emphasizes the importance of consulting a surgeon for personalized medical advice	GPT-3.5 has the potential to enhance patient education during rhinoplasty recovery but should not replace personalized advice from qualified healthcare professionals
24. Coskun et al., 2023; Turkey[46]	Quantitative	Urology; GPT-3.5's utility in providing patient information on	GPT-3.5; patients	59 questions were derived from the EAUPI website; response evaluation by 2 urologists for content	GPT-3.5's responses were suboptimal in accuracy and quality, with an average F1 score: 0.426,	Caution should be exercised when using GPT-3.5 for

		prostate cancer compared to the European Association of Urology Patient Information (EAUPI)		accuracy/similarity and quality using precision, recall, F1 score, cosine similarity, and the General Quality Score (GQS)	precision: 0.349, recall: 0.549, cosine similarity: 0.609, and GQS: 3.62□±□0.49; no answer achieved the maximum GQS of 5	patient information on prostate cancer
25. Durairaj et al., 2023; USA, Italy[47]	Quantitative	Plastic Surgery; to compare GPT-3.5's performance in responding to patient questions on septorhinoplasty to the responses from a rhinoplasty surgeon	GPT-3.5; patients	6 hypothetical questions on septorhinoplasty were designed by the author; blinded responses of a board-certified rhinoplasty surgeon and GPT-3.5 were evaluated by 7 rhinoplasty surgeons for empathy, accuracy, completeness, overall quality, preferred response	GPT-3.5 outperformed the surgeon response in accuracy, completeness, and overall quality; empathy rating did not significantly differ; GPT-3.5 responses were preferred in 81% of cases	GPT-3.5 has the potential to assist surgeons in educating and counseling patients on septorhinoplasty
26. Kianian et al., 2023; USA[48]	Quantitative	Ophthalmology; GPT-3.5's and Bard's ability to produce patient-targeted health information on uveitis and to improve the readability of online health information	GPT-3.5, Bard; patients	2 prompts for generating patient-focused health information about uveitis; 9 patient-focused uveitis web page texts from the first Google page were asked to be rewritten for readability; responses were analyzed for readability using the Flesch-Kincaid Grade Level (FKGL); appropriateness rated by 2 fellowship-trained uveitis specialists	Appropriateness GPT-3.5: 100%/Bard: 88.9%; GPT-3.5 provided significantly more comprehensible responses (mean FKGL: 6.3) compared to Bard (mean FKGL: 10.5); online uveitis health information averaged a FKGL of 11.0, GPT-3.5 had a mean FKGL of 8.0, Bard had a mean FKGL of 11.1	GPT-3.5 outperforms Bard in generating/rewriting patient-friendly health information on uveitis
27. Seth et al., 2023; Australia[49]	Quantitative	Hand Surgery; GPT-3.5's precision and comprehensiveness of answers on the management of carpal tunnel syndrome (CTS); to assess the safety of GPT-3.5's medical advice	GPT-3.5; patients/caregivers	2 plastic surgeons developed 6 CTS questions and evaluated responses for accuracy, reliability, comprehensiveness, and reference generation; simulated doctor-patient interactions were employed to assess the safety of GPT-3.5's medical advice	GPT-3.5 provided relevant but superficial information on CTS; references were considered insufficient; during the simulated doctor-patient interactions, GPT-3.5 suggested a diagnostic pathway that differed from the widely accepted clinical consensus on CTS diagnosis	GPT-3.5 has the potential to provide general medical information to patients but requires refinement, particularly regarding accurate referencing and depth of information
28. Inojosa et al., 2023; Germany[50]	Quantitative	Neurology; GPT-4's performance in communicating medical information relevant to multiple sclerosis (MS) to medical professionals and MS patients	GPT-4; patients/caregivers	64 clinical scenarios related to MS treatment were manually created and used to generate one explanation each for general practitioners and MS patients; response grading by 3 medical doctors specialized in MS treatment for humanness, accuracy, reliability, writing quality; readability assessment using the Flesch-Kincaid Grade Level (FKGL)	Median humanness score (on a 5-point Likert scale): 5; median correctness score: 4.25; median relevance score: 4; mean FKGL of 15.26 for general practitioners' and 63.14 for MS patients' information	GPT-4 shows promise for communicating medical information related to MS; validation and correction by expert care providers are necessary to ensure patient safety
29. Lyons et al., 2023; USA[51]	Quantitative	Ophthalmology; to evaluate the triage performance of AI chatbots for ophthalmic conditions	GPT-4, Bing Chat; patients/caregivers	44 clinical vignettes were developed based on a literature review of common emergency room ophthalmologic diagnoses; comparison of performance with WebMD Symptom Checker and 8 ophthalmology trainee respondents; response evaluation by 2 experts for accurate diagnosis listed in the top 3 possible diagnoses and correct triage urgency, grossly inaccurate statements, mean reading grade level, mean response word count, proportion with attribution, and most common sources cited	Ophthalmology trainees achieved the highest correct diagnosis rate (95%), followed by GPT-4 (93%), Bing Chat (77%), and WebMD Symptom Checker (33%); GPT-4 scored highest in triage accuracy (98%), followed by ophthalmology trainees (86%) and Bing Chat (84%); gross inaccuracies were found in 0% of responses by GPT-4 and trainees, 14% by Bing Chat, and 50% by WebMD Symptom Checker	GPT-4 offers high diagnostic and triage accuracy for ophthalmic conditions comparable to that of ophthalmology trainees, suggesting potential utility as a triage tool in healthcare settings
30. Babayiğit et al., 2023; Turkey[52]	Quantitative	Periodontology; GPT-3.5's ability to answer the most frequently asked questions on different topics in periodontology	GPT-3.5; patients	70 most-frequently asked patient questions generated by GPT-3.5 on 7 different periodontology topics determined by periodontists; 20 periodontists were contacted via email to evaluate the answers for accuracy and completeness	Mean accuracy score (7-point Likert scale): 5.5; mean completeness score (3-point Likert scale): 2.34; statistically significant differences in performance between subjects	GPT-3.5 can be an informational resource for patients and periodontists, but expert supervision is needed to address potential inaccuracies
31. Mondal et al., 2023; India[53]	Quantitative	Discipline not specified; GPT-3.5's ability to answer patient questions related to lifestyle-related diseases and disorders	GPT-3.5; patients	20 fictional cases with 4 lifestyle-related disease questions each were created; content validity checked by a public health expert; response evaluation for accuracy, guidance, sentiment analysis, readability, and content evaluation by two primary care physicians	Average accuracy score (3-point assessment scale from 0 to 2): 1.83; average guidance score: 1.9; high Flesh-Kincaid Grade Level of 14.37 and Flesch Reading Ease Score of 27.8; responses were in a natural and positive tone	GPT-3.5 provides accurate responses and adequate guidance for lifestyle-related health diseases and disorders

32. Kim et al., 2023; South Korea[54]	Quantitative	Neurology; GPT-3.5's versus GPT-4's performance in providing educational information on epilepsy	GPT-3.5/GPT-4; patients	57 epilepsy questions were developed based on the Korean Epilepsy Society's 'Epilepsy Patient and Caregiver Guide' and prompted twice; response evaluation by 2 epileptologists for educational value/correctness	70% of GPT-4's responses had sufficient educational value; 28% were correct but inadequate; no response was entirely incorrect; GPT-4 outperformed GPT-3.5 and was often on par or better than the actual guide	GPT-4 can be a valuable tool in delivering reliable epilepsy-related information
33. Song et al., 2023; China[55]	Quantitative	Urology; effectiveness of LLMs in providing medical consultations and patient education on urolithiasis	Bard, Claude, GPT-4, Bing Chat; patients/caregivers	21 questions from online consultation platforms, surveys conducted among hospitalized urolithiasis patients, and researchers' clinical experience; 2 case scenarios with different complexity; response evaluation by 3 urolithiasis experts for accuracy, ease of understanding, comprehensibility, human caring	Claude consistently scored the highest in all dimensions; GPT-4 ranked second in accuracy, with shortcomings in empathy and human caring; Bard had the lowest accuracy and overall performance	Claude shows superior performance compared to the other 3 LLMs in providing consultations and education on urolithiasis
34. Bitar et al., 2022; Saudi Arabia, USA[24]	Quantitative	Gynecology; to assess if BERT text summarization increases women's knowledge about HPV	BERT; patients	386 women aged ≥ 20 years recruited via Amazon Mechanical Turk were randomly assigned to 2 groups: 1. BERT summarized text, 2. original text on HPV based on 3 publications; a 29-item questionnaire based on Waller et al.'s HPV knowledge measure was used to assess participants' pre- and post-knowledge	Women who read the original texts were more likely to correctly answer 2 questions on the general HPV knowledge subscale and 1 question on the HPV testing knowledge subscale; HPV vaccination knowledge did not significantly differ	BERT text summarization could be a valuable tool in public health education, providing a balance between information completeness and reader time efficiency
35. Zalzal et al., 2023; USA[56]	Quantitative	Otolaryngology; GPT-3.5's utility for answering otolaryngology-related questions from the lay public	GPT-3.5; patients	30 commonly asked questions by patients/families were collected over 3 months by pediatric otolaryngologists; response ratings by 2 board-certified otolaryngologists and 13 lay public graders for correctness or confidence of accuracy	Experts: 98.3% of questions correct; non-experts: 79.8% confidence in GPT-3.5's response accuracy	GPT-3.5 can serve as a helpful medical information tool; while physicians rate its information as accurate and comprehensive, laypersons lack confidence in GPT-3.5
36. Chervenak et al., 2023; USA[57]	Quantitative	Gynecology; GPT-3.5's performance in responding to fertility-related questions	GPT-3.5; patients	1. 17 infertility questions from the FAQ of the Centers for Disease Control (CDC), response assessment by 2 physicians for sentiment analysis, factual statements, incorrectness, and references; 2. 2 validated fertility knowledge surveys, evaluation of percentiles compared to published population data; 3. 7 statements from the American Society for Reproductive Medicine Committee converted into questions, assessed for identification of missing facts	1. GPT-3.5 responses matched CDC's in length, factual content, sentiment, and subjectivity; 6.12% of GPT-3.5's factual statements were incorrect, only one (0.7%) provided a reference; 2. GPT-3.5 scored at the 87th percentile for the Cardiff Fertility Knowledge Scale and at the 95th percentile for the Fertility and Infertility Treatment Knowledge Score; 3. all 7 missing facts for the summary statements were reproduced	GPT-3.5 produces relevant and meaningful responses to fertility-related questions comparable to established resources
37. Bushuven et al., 2023; Germany[58]	Quantitative	Emergency Medicine; GPT-3.5's and GPT-4's performance in supporting parents in Basic Life Support (BLS) and Pediatric Advanced Life Support (PALS)	GPT-3.5/GPT-4; patients	22 case vignettes describing prototypical BLS (n=2)/PALS emergencies (n=20), developed and validated by 5 emergency physicians and prompted 3 times; response evaluation for diagnostic accuracy, emergency call advice, and validity of advice	GPT-3.5/GPT-4 accurately diagnosed the condition in 94% of cases, advised calling emergency services in 54% of cases, provided correct first aid instructions in 45% of cases, and incorrectly recommended advanced life support techniques in 13.6% of cases	The reliability and safety of GPT-3.5/GPT-4 as emergency support tools are questionable, but they show potential for aiding in diagnosing pediatric emergencies
38. Jeblick et al., 2023; Germany[59]	Quantitative	Radiology; to assess the quality of GPT-3.5 in generating simplified radiology reports	GPT-3.5; patients	Three fictitious radiology reports were created by a radiologist and simplified by GPT-3.5 15 times each; quality assessment by 15 radiologists for actual correctness, completeness, and potential harm	Factual correctness: 75% "Agree/Strongly agree"; incorrect passages in 51% of reports; missing relevant information in 22% of reports; potentially harmful content in 36% of reports	GPT-3.5 shows potential in simplifying radiology reports but needs refinement to ensure accuracy and prevent harm
39. Samaan et al.; 2023; USA[60]	Quantitative	General Surgery; GPT-3.5's accuracy and reproducibility in answering patient questions about bariatric surgery	GPT-3.5; patients	151 questions from professional societies, health institutions, and Facebook support groups prompted twice each; response grading for accuracy and reproducibility by 2 board-certified bariatric surgeons	Comprehensive: 86.8% of responses; reproducible: 90.7% of responses	GPT-3.5 is a useful information resource for patients about bariatric surgery, but it should complement, not replace, standard care from healthcare professionals

40. Zhou et al.; 2023; Germany, India, Spain[61]	Quantitative	Gastroenterology; GPT-3.5's and GPT-4's potential in disseminating gastric cancer knowledge, providing consultation recommendations, and interpreting gastroscopy reports	GPT-3.5/GPT-4; patients/caregivers	23 gastric cancer questions prompted 3 times, evaluation of appropriateness and consistency of responses; case materials from the Chinese Medical Case Repository, Journal of Medical Case Reports, and F1000 Research prompted 3 times to assess consultation recommendations and endoscopy report analysis for abnormalities and consistency	GPT-4 outperformed GPT-3.5 in all tasks; gastric cancer questions: 91.3% appropriate, 95.7% consistent (GPT-3.5: 78.3%/82.6%); consultation recommendations: 80.4% appropriate, 82.6% consistent (GPT-3.5: 69.6%/73.9%); endoscopy report analysis: 69.6% appropriate, 65.2% consistent (GPT-3.5: 56.5%/58.7%)	GPT-4 shows potential in disseminating medical knowledge and assisting in medical consultation but should not be a substitute for professional medical advice
41. Oniani et al., 2023; USA[62]	Quantitative	Discipline not specified; Effectiveness of neural machine translation (NMT) models in translating health illiterate language in patient education materials	BERT, BioBERT, BioClinicalBERT; patients	A corpus for training NMT models was created using data on patient education materials from MedlinePlus.gov, Drugs.com, Mayoclinic.org, and Reddit.com; conversion into illiterate language using the CDC Plain Language Thesaurus; evaluation of NMT models using BLEU score for translation quality	BiLSTM outperformed LLMs with a mean BLEU score of 41.578, followed by BERT: 33.582, BioBERT: 33.278, and BioClinicalBERT: 31.191	NMT models show effectiveness in translating health illiterate language into patient education materials
42. Hernandez et al., 2023; Barbados, USA[63]	Quantitative	Endocrinology; GPT-3.5's correctness and consistency in responding to questions on type 2 diabetes mellitus (T2DM) and associated complications	GPT-3.5; patients	70 questions about T2DM and its complications were developed by physicians and evaluated by research staff; prompted 3 times each; response evaluation by 2 internal medicine board-certified physicians	98.5% of responses were appropriate; 1.4% of responses were inappropriate but still met minimal standards of care	GPT-3.5 demonstrates potential as a supplementary tool for diabetes education
43. Kuşcu et al., 2023; Turkey, Iran[64]	Quantitative	Otolaryngology; GPT-4's accuracy and reliability in responding to head and neck cancer (HNC) questions	GPT-4; patients/caregivers	154 HNC questions from professional societies, institutions, patient support groups, and social media; response grading by 2 head and neck surgeons for accuracy and reproducibility	Comprehensive/correct: 86.4% of responses; reproducible: 94.1% of responses; no responses completely inaccurate/irrelevant	GPT-4 has the potential to serve as a valuable information source on HNC for patients and healthcare professionals
44. Biswas et al., 2023; UK[65]	Quantitative	Ophthalmology; GPT-3.5's accuracy and information quality in answering myopia questions	GPT-3.5; patients	11 questions on myopia were constructed based on the "frequently asked questions on myopia" webpage of the Association of British Dispensing Opticians; prompted 5 times each; response evaluation by 5-members of the optometry teaching and research staff for accuracy and response quality	24% of questions were rated as very good, 49% as good, 22% as acceptable, 3.6% as poor, and 1.8% as very poor; information quality was good for 90.9% of questions and acceptable for one question (9.1%)	GPT-3.5 shows potential in providing information on myopia, but limitations and inaccuracies need to be addressed before its implementation in clinical settings
45. Chiesa-Estomba et al., 2023; Spain, Austria, Belgium, France, Italy[66]	Quantitative	Otorhinolaryngology; to assess the level of agreement between GPT-3.5 and expert sialendoscopists (EESS) in clinical decision-making and patient information support for the management of salivary gland disorders	GPT-3.5; patients	6 questions based on the most common clinical scenarios in 3 sialendoscopy clinics; 10 EESS responded via e-questionnaire and were compared against GPT-3-5 by another set of 10 EESS, assessing the level of agreement	Mean agreement score (5-point Likert scale): GPT-3.5: 3.4, ES: 4.1; mean therapeutic alternatives number: GPT-3.5: 3.3, ES: 2.6	GPT-3.5 is a promising tool in the clinical decision-making process within the salivary gland clinic
46. Decker et al., 2023; USA[67]	Quantitative	General Surgery; to compare GPT-3.5's readability, accuracy, and completeness with surgeon-generated information on the risks, benefits, and alternatives (RBAs) of common surgical procedures	GPT-3.5; patients/caregivers	6 RBAs for common surgical procedures were generated using GPT-3.5 by a multidisciplinary group of surgeons and compared against 5 surgeon-generated RBAs for each of the 6 surgical procedures for readability, accuracy, and completeness using a rubric with recommendations from LeapFrog, the Joint Commission, and the American College of Surgeons by at least 2 blinded reviewers	Mean composite scores for completeness and accuracy: GPT-3.5: 2.2, surgeons: 1.6; mean readability scores: GPT-3.5: 12.9, surgeons: 15.7	GPT-3.5 has the potential to enhance informed consent documentation by providing more readable, accurate, and complete information compared to surgeon-generated content
47. Kaarre et al., 2023; USA, Sweden[68]	Quantitative	Orthopedics; GPT-4's usefulness in answering questions by patients and non-orthopedic medical doctors on anterior cruciate ligament (ACL) surgery	GPT-4; patients	20 questions on ACL surgery developed based on a literature search and frequently asked patient questions for patients and non-orthopedic doctors; response evaluation by 4 orthopedic sports medicine surgeons for correctness, completeness, and adaptiveness	GPT-4 patient responses fully or majority correct: 65%; mean correctness score (2-point scale from 0 to 2): patients: 1.69, surgeons: 1.66; mean completeness score: patients: 1.51, surgeons: 1.64; mean adaptiveness score: patients: 1.75, surgeons: 1.73	GPT-4 has the potential as a supplementary tool for patient education on ACL surgery but cannot replace the expertise of orthopedic sports medicine surgeons

48. Ferreira et al., 2023; USA[69]	Quantitative	Dermatology; GPT-4's appropriateness in responding to common questions by dermatology patients	GPT-4; patients	31 questions on 6 common skin conditions and queries were created by 3 dermatologists based on their experience and a literature review and prompted 3 times; response grading as "appropriate" or "inappropriate" by 3 dermatologists	88% of responses were appropriate; 12% of responses were inappropriate; 16.1% of responses had an inappropriate response average, with a minimum of 2 dermatologists rating 2 out of 3 responses as inappropriate	GPT-4 shows potential as a public dermatology resource, but it should not replace professional medical advice and remain a supplementary informational tool
49. Truhn et al., 2023; Germany[70]	Quantitative	Radiology, Orthopedics; GPT-4's validity of patient treatment recommendations for common knee and shoulder orthopedic conditions using clinical MRI reports	GPT-4; patients	20 anonymized reports out of 94 knee and 38 shoulder MRI studies were selected by a musculoskeletal radiologist and prompted twice for German-English translation and the provision of treatment recommendations; response evaluation by 2 orthopedic surgeons for overall quality, scientific and clinical basis, and clinical usefulness and relevance	Quality of treatment recommendations was rated as good (10%), very good (60%), or excellent (30%) for the knee and shoulder; recommendations were mainly up-to-date and consistent, adhering to clinical and scientific evidence; no signs of hallucinations or nonsensical responses, but a tendency to provide generic and unspecific answers	GPT-4 shows promise in offering accurate and clinically relevant treatment recommendations for orthopedic knee and shoulder issues but should not replace consultations with specialists for treatment advice
50. Hurley et al., 2023; USA[71]	Quantitative	Orthopedics; GPT-3.5's quality and readability of information regarding shoulder stabilization surgery	GPT-3.5; patients	23 patient questions on shoulder stabilization surgery were developed based on prior studies; response evaluation by 3 residents for quality (DISCERN score, JAMA benchmark criteria) and readability (Flesch-Kincaid Reading Ease Score (FRES) and grade level (FKGL))	JAMA benchmark criteria score: 0 (no reference cited); DISCERN score: 60 (considered good); FRES: 26.2; FKGL of a college graduate	GPT-3.5 has the potential to provide high-quality answers to questions relating to shoulder stabilization surgery, but it is unclear where the answers originated
51. Cankurtaran et al., 2023; Turkey[72]	Quantitative	Gastroenterology; GPT-3.5's performance in answering inflammatory bowel disease (IBD) questions for patients and healthcare professionals	GPT-3.5; patients/caregivers	5 patient questions each on Crohn's disease and Colitis Ulcerosa based on Google Trends; 5 questions each generated by 4 gastroenterologists; response evaluation for reliability and usefulness by 2 gastroenterologists	Mean patient reliability/usefulness scores (7-point Likert scale): professional sources 5/5.2, patient-derived responses: 4/4.35	GPT-3.5 exhibits partial reliability and usefulness in the context of IBD but has limitations and deficiencies
52. Birkun et al., 2023; Russia, India[73]	Quantitative	Emergency Medicine; to evaluate the performance of LLMs in providing guideline-consistent advice on help to a non-breathing victim in emergencies	Bard, Bing Chat; patients	Bing Chat and Bard were prompted 20 times each with the query "What to do if someone is not breathing?"; original and self-corrected response rated by 2 authors for compliance with the Resuscitation Council UK Guidelines	LLM's responses lacked guideline-consistent instructions for helping a non-breathing victim; Bing Chat: 9.5% compliance; Bard: 11.4% compliance; LLMs overestimated the quality of their response compared to expert ratings (10-point Likert scale); Bing Chat: 7 points; Bard: 9.0 points; LLMs denied containing guidelines-inconsistent instructions	Bing Chat and Bard provide understandable but unreliable resuscitation information, lacking essential details and occasionally including harmful directives
53. Pushpanathan et al., 2023; Singapore, China[74]	Quantitative	Ophthalmology; proficiency of LLMs in addressing queries related to ocular symptoms	GPT-3.5/GPT-4, Bard; patients	37 questions on ocular symptoms were developed by a team of 5 ophthalmologists and clinical optometrists considering common online health information sites; random presentation to 3 ophthalmologists and evaluation for accuracy, comprehensiveness, evaluation of self-awareness levels through prompts for self-correction; qualitative analysis of poorly rated responses by two ophthalmologists	GPT-4 exhibited higher average total accuracy (8.2 of 9) compared to GPT-3.5 (7.5) and Google Bard (7); comprehensiveness was good without significant differences; GPT-3.5 issued a general disclaimer when prompted to self-check, emphasizing the need for additional personal medical information; GPT-4 and Google Bard consistently asserted the accuracy of their original responses, even when deemed as 'poor' or 'borderline'	GPT-4 demonstrates superior performance in addressing ophthalmologic queries, highlighting its utility in providing accurate and comprehensive responses
54. Shao et al., 2023; China[75]	Quantitative	Thoracic Surgery; GPT-3.5's appropriateness and comprehensiveness for perioperative patient education in thoracic surgery in English and Chinese contexts	GPT-3.5; patients	37 questions focused on perioperative thoracic surgery patient education based on guideline-based topics and personal experience prompted in English and Chinese; response evaluation by 35 reviewers with thoracic surgery experience for appropriateness and comprehensiveness	92% of responses were rated as "qualified" both in English and Chinese contexts, 8% of responses in both languages were rated as "unqualified"	GPT-3.5 shows promise in providing appropriate and comprehensive information, which could enhance patient education and clinical service quality in thoracic surgery
55. Vaira et al., 2023; Italy, Belgium, Spain, France[76]	Quantitative	Head and Neck Surgery; GPT-4's accuracy in answering questions and solving clinical scenarios related to head and neck surgery	GPT-4; patients/caregivers	144 questions (50% each open-ended/binary) and 15 clinical scenarios developed by 18 head and neck surgeons from 14 Italian centers; response evaluation for accuracy and completeness by the same 18 surgeons; reference assessment by 2 reviewers; comparison of performance with a resident surgeon	84.7% correct response rate for closed-ended questions; median accuracy score of 6 (of 6) for open-ended questions; median completeness score of 3 (of 3) for open-ended questions; fully or nearly fully correct diagnoses in 81.7% of clinical scenarios; complete diagnostic or therapeutic	GPT-4 demonstrates a good level of accuracy in responding to head and neck surgery but should not be considered a reliable support for decision-making in clinical settings

				by 3 reviewers	procedures proposed in 56.7% of cases; the resident significantly outperformed GPT-4 in all domains	
56. Chen et al., 2023; USA[77]	Quantitative	Oncology; GPT-3.5's performance in providing cancer treatment recommendations concordant with National Comprehensive Cancer Network (NCCN) guidelines	GPT-3.5; patients	4 zero-shot prompts for treatment recommendations for each of 26 cancer diagnosis descriptions; 3 board-certified oncologists assessed concordance with the 2021 NCCN guideline based on 5 scoring criteria	100% of outputs included at least one NCCN-concordant treatment; 34.3% of outputs also recommended one or more non-concordant treatments; hallucinated treatments in 13 of 104 (12.5%) outputs	One-third of GPT-3.5's treatment recommendations were at least partially non-concordant with NCCN guidelines; therefore, clinicians should advise patients that chatbots are not reliable sources of treatment information
57. Bellinger et al., 2023; USA[78]	Quantitative	Otolaryngology; GPT-3.5's performance in responding to questions on Benign Paroxysmal Positional Vertigo (BPPV) compared to Google webpages	GPT-3.5; patients	5 questions based on the top 30 Google search results for BBPV prompted 3 times; response assessment and comparison to Google Websites addressing these questions by 2 first-year medical student reviewers for readability (Flesch□Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRE)), quality (DISCERN score), and understandability and actionability (PEMAT-P); accuracy and currency evaluation by 2 physicians	Experts rated GPT-3.5 information as accurate (mean score 4.2 of 5) and current (mean score 4.3); GPT-3.5 had higher FKGL scores (mean 13.9) compared to Google search results (mean 10.7); Google webpages had higher total DISCERN scores (mean 56.5) compared to GPT-3.5 individual responses (mean 17.5); Google webpages scored higher in understandability (82.3%) compared to GPT-3.5 individual responses (72.3%)	GPT-3.5 provides accessible information but is generally of lower quality, readability, understandability, and actionability compared to Google webpage resources
58. Nielsen et al., 2023; Denmark[79]	Quantitative	Otorhinolaryngology; GPT-4's accuracy in providing relevant medical information on otolaryngology (ORL) questions	GPT-4; patients	27 questions on 9 ORL conditions; response evaluation by 13 physicians from a tertiary ORL department for accuracy, relevance, and depth of responses; GPT-4 was asked to evaluate the responses on the same metrics as an otolaryngologist	Mean score of 3.4 (of 5) for accuracy, relevance, and depth; highest rating for relevance (mean score of 3.7); lowest rating for depth (mean score of 3); self-assessment rating of GPT-4 was 5 for all categories	GPT-4 demonstrates promise in providing relevant and accurate medical information but requires enhancements in response depth and mitigating potential biases
59. Sezgin et al., 2023; USA[80]	Quantitative	Gynecology; LLMs' quality of responses compared to Google Search results in addressing questions about postpartum depression (PPD)	GPT-4, Bard; patients	14 PPD-related patient questions from the American College of Obstetricians and Gynecologists; response evaluation by 2 board-certified physicians using the GRADE-informed scale	Mean GRADE score for GPT-4: 3.9; significant higher quality compared to Bard (Z=2.143; adjusted P=.048) and Google Search (Z=3.464; adjusted P<.001)	GPT-4 and Bard exhibit potential in delivering clinically accurate responses that are of higher quality compared to those obtained from Google Search
60. Floyd et al., 2023; USA[81]	Quantitative	Radiation Oncology; GPT-3.5's/GPT-4's accuracy and comprehensiveness in answering patient questions related to radiation oncology	GPT-3.5/GPT-4; patients/caregivers	252 patient-centered questions, including 28 templates applied to 9 cancer types based on the Patient Concerns Inventory for Head and Neck Cancer and the National Cancer Institute's website "Questions to Ask Your Doctor About Treatment"; evaluation by 2 independent radiation oncology residents for accuracy and comprehensiveness	34.1% of answers contained inaccurate information, 26.2% contained correct information but missed essential context, 39.7% of responses were correct and comprehensive; among inaccurate responses, 71% were categorized as potentially appropriate within a context other than the prompt, and 29% were deemed inaccurate in any clinical context; GPT-4 performed similarly on a subset of questions	GPT-3.5/4 fail to consistently generate accurate and comprehensive responses to the majority of radiation oncology patient-centered questions
61. Uz et al., 2023; Turkey[82]	Quantitative	Rheumatology; GPT-3.5's reliability and usefulness in providing information about common rheumatic diseases	GPT-3.5; patients/caregivers	7 common rheumatic diseases were identified using the American College of Rheumatology and European League against Rheumatism guidelines; Google Trends was used to determine the 4 most frequently searched keywords for each disease; keywords were used by different users in an ongoing chat; response evaluation for reliability and usefulness by 2 physical medicine and rehabilitation specialists	Highest reliability score for osteoarthritis (mean: 5.62 of 7); highest usefulness score for ankylosing spondylitis (mean: 5.87 of 7); no significant difference between the subjects in terms of reliability and usefulness	Although GPT-3.5 is reliable and useful for patients, it may contain false answers
62. Athavale et al., 2023; USA[83]	Quantitative	Vascular Surgery; to assess the potential of chatbots in answering chronic venous disease patient	GPT-3.5/GPT-4, LLaMA (Clinical Camel);	2 questionnaires (1 for non-complex medical/administrative matters, 1 for complex medical chronic venous disease questions based on	Non-complex medical questions: GPT-4: 100% appropriate and complete, GPT-3.5: 70%; complex medical questions: GPT-4: 75% appropriate and	GPT-4 demonstrates potential in responding to administrative/non-complex

		questions and electronic health record inbox management	patients/caregivers	patient messages) consisting of 20 questions each; complex medical questions were also posed to Clinical Camel; response grading by 2 physicians for appropriateness and completeness	complete, GPT-3.5: 45%; Clinical Camel: 0% appropriate and complete, 25% appropriate but incomplete, 25% neither appropriate nor inappropriate, 50% wrong	medical and complex medical questions related to chronic venous disease
63. Li et al., 2023; USA, China[84]	Quantitative	Discipline not specified; to create a specialized LLM with enhanced accuracy in medical advice	GPT-3.5, LLaMA (ChatDoctor); patients/caregivers	Dataset curation using roughly 100,000 interactions from HealthCareMagic and 10,000 conversations from iCliniq; MedlinePlus/Wikipedia as an external knowledge brain; performance evaluation by testing various contemporary medical queries; BERTScore was employed to compute precision, recall, and F1 scores for ChatDoctor and GPT-3.5 on questions from the iCliniq database, with responses from physicians used as benchmark	ChatDoctor significantly outperformed GPT-3.5 in precision, recall, and F1 scores and provided more sufficient and reliable answers on novel diseases and drugs	ChatDoctor has the potential to improve accuracy and efficiency in medical diagnosis, reducing the workload for medical professionals
64. Seth et al., 2023; Australia[85]	Quantitative	Plastic Surgery; to assess the efficacy of employing LLMs in obtaining and synthesizing information about rhinoplasty	GPT-3.5, Bard, Bing Chat; patients	6 questions on rhinoplasty were developed by 3 board-certified plastic surgeons and prompted to GPT-3.5, Bard, and Bing Chat; response evaluation by comparing them with current healthcare guidelines for rhinoplasty and through evaluation by the panel of plastic surgeons; readability assessment using the Flesch Reading Ease Score (FRES), Flesch–Kincaid Grade Level (FKGL), and Coleman–Liau Index (CLI); suitability assessment with modified DISCERN score	Bard and GPT-3.5 showed a significantly higher mean FRES than Bing Chat: Bard: 47.47, GPT-3.5: 37.68, Bing Chat: 18.29; Bard and GPT-3.5 had a lower mean FKGS: Bard: 9.7, GPT-3.5: 10.15, Bing Chat: 18.25; Bard and GPT-3.5 exhibited a lower CLI: Bard: 10.83, GPT-3.5: 12.17, Bing Chat: 12.00; Bard had the highest DISCERN score: Bard: 46.33, GPT-3.5: 42.17, Bing Chat: 35.75	The use of LLMs such as GPT-3.5, Bard, and Bing Chat to obtain detailed information about specific surgical procedures such as rhinoplasty demonstrate potential, but challenges regarding their depth and specificity remain.
65. Kuckelman et al., 2024; USA[110]	Quantitative	Radiology; Bing Chat's accuracy in providing patient education for common radiologic exams	Bing Chat; patients	10 questions each for MRI Spine, CT abdomen, and bone biopsy were developed by the authors and Bing Chat itself (50% each); prompted twice using 3 different chatbot settings; response grading for accuracy and completeness by 2 independent reviewers compared to radiologyinfo.org	93% of responses "entirely correct", 7% "mostly correct"; 65% of responses "complete", 35% of responses "mostly complete"	Bing Chat offers precise responses regarding radiology exams and procedures, indicating its potential to enhance the patient experience in radiology
66. Lockie et al., 2023; Australia[86]	Quantitative	General Surgery; to evaluate a GPT-3.5 generated patient information leaflet (PIL) against a surgeon-generated version	GPT-3.5; patients	28 patients undergoing laparoscopic cholecystectomy and 16 doctors from 2 hospitals in Melbourne were asked to complete a survey based on a validated evaluation instrument for PIL; comparison of the PIL about laparoscopic cholecystectomy generated by GPT-3.5 versus one developed by surgeons	Patients scored GPT-3.5 and surgeon-generated PILs similarly (both median: 8 (of 8)); doctors rated GPT-3.5's PIL slightly higher (median: 7 for GPT-3.5 versus 6 for surgeons)	GPT-3.5's-generated PIL was comparable to or slightly better than the surgeon-generated version, indicating the feasibility of using LLMs for PIL creation
67. Haver et al., 2023; USA[87]	Quantitative	Radiation Oncology; to evaluate the effectiveness of LLMs in simplifying responses to patient questions on lung cancer and lung cancer screening (LCS) to improve readability and clinical appropriateness	GPT-3.5/GPT-4, Bard; patients	19 questions, each prompted 3 times, about lung cancer and LCS were posed to GPT-3.5 to generate baseline responses; assessment for readability and accuracy by 3 fellowship-trained cardiothoracic radiologists; simplified baseline responses evaluated by the same radiologists	Baseline: GPT-3.5 Flesch reading ease score: 49.7, Flesch-Kincaid readability grade: 12.6; Simplified: GPT-3.5 reading ease: 62, readability grade: 10, GPT-4 reading ease: 68, readability grade: 9.6, Bard reading ease: 71, readability grade: 8.2; responses were deemed clinically accurate in 84% (GPT-3.5), 79% (GPT-4), and 95% (Bard)	GPT-3.5, GPT-4, and Bard demonstrate potential in generating and simplifying responses to patient questions on lung cancer and LCS
68. Li et al., 2023; USA[88]	Quantitative	Radiology; GPT-3.5's potential in simplifying radiology reports to the reading level of the average United States adult	GPT-3.5; patients	400 deidentified radiology reports (100 each for X-ray (XR), ultrasound (US), MRI, CT) were randomly selected from the institution's database and prompted for simplification; response evaluation for report length, Flesch reading ease score (FRES), and Flesch-Kincaid reading level (FKRL)	Following simplification, all reports had an FKRL <8.5, with a mean increase in FRES of 46 points and a mean decrease in FKRL by 5-grade levels to an average of <6th-grade level	GPT-3.5 effectively simplifies radiology reports to the reading level of the average United States adult, but further evaluation of accuracy and impact on patient comprehension is needed
69. Scheschenja et al., 2023; Germany[89]	Quantitative	Radiology; GPT-3.5's and GPT-4's accuracy in providing patient education regarding specific interventional radiology (IR) procedures	GPT-3.5/GPT-4; patients	133 questions on port implantation, percutaneous transluminal angioplasty, and transarterial chemoembolization procedures developed by 2 radiology residents and validated by a third radiologist; response evaluation by 2 radiologists for accuracy	GPT-3.5: 30.8% "completely correct", 48.1% "very good"; GPT-4: 35.3% "completely correct", 47.4% "very good"; GPT-4 was significantly more accurate than GPT-3.5; no responses were identified as potentially harmful	GPT-3.5 and GPT-4 show potential for safe and relatively accurate in-depth patient education in IR procedures

70. Gordon et al., 2023; USA[90]	Quantitative	Radiology; GPT-4's accuracy, relevance, and readability in answering patients' imaging-related questions and evaluating the influence of a patient-directed prompt on these parameters	GPT-4; patients	22 questions developed based on the expertise of 4 radiologists and existing literature prompted 3 times by using an unstructured and structured prompt; response evaluation for accuracy, relevance, consistency, and readability (Flesch–Kincaid Grade Level (FKGL)) by the 4 radiologists and patients	Accuracy: 83% unprompted, 87% structured prompt; relevance: 98.5% unprompted, 98.8% structured prompt; FKGL: 13.6 unprompted, 13.0 structured prompt; consistency: 72% unprompted, 86% structured prompt; patient utility assessment: 92%-97% responses deemed as relevant and helpful by patients	GPT-4 has the potential to provide accurate and relevant answers to patient-centered imaging questions but is cautioned against immediate clinical implementation due to imperfections in accuracy, consistency, and readability.
71. Stroop et al., 2023; Germany[91]	Quantitative	Neurosurgery; GPT-3.5's accuracy in providing medical information on acute lumbar disc herniation (LDH)	GPT-3.5; patients	52 spinal surgeons completed an online survey, imagining themselves as patients with acute LDH, and interacted with GPT-3.5 for information; quality evaluation of responses based on predefined categories by the spinal surgeons; responses were compared to a standardized informed consent sheet for LDH	97% of GPT-3.5's responses understandable; 55% of responses medically comprehensive; GPT-3.5 covered 48% informed consent form information for LDH	GPT-3.5 shows potential in supporting medical communication with patients by providing understandable responses related to LDH
72. Coraci et al., 2023; Italy, Bulgaria[92]	Quantitative	Orthopedics; effectiveness of a GPT-3.5-generated questionnaire for assessing low back pain (LBP) compared to routinely used and validated questionnaires	GPT-3.5; patients	GPT-3.5 was prompted to generate a questionnaire for the assessment of the LBP (ChatQ) that consisted of 10 questions in Italian; ChatQ was administered to 20 Italian-speaking patients with a history of LBP for self-compilation compared to the Numeric Pain Rating Scale (NRS) and 3 validated questionnaires for back pain: Oswestry Disability Index (ODI), Quebec Back Pain Disability Scale (QBPDS), and Roland-Morris Disability Questionnaire (RMDQ)	ChatQ: median score: 8/17; ODI: median score: 12%; QBPDS: median score: 9/100; RMDQ: median score: 3/24; NRS: median score: 4/10; strong correlation between ODI and ChatQ, moderate correlation between QBPDS and ChatQ, no statistical correlation between ChatQ and RMDQ or NRS	ChatQ, generated by GPT-3.5, offers potential benefits for the assessment of LBP but cannot replace established validated questionnaires
73. Ye et al., 2023; Canada, Germany[93]	Quantitative	Rheumatology; GPT-4's quality of responses compared to those of rheumatologists for real rheumatology patient questions	GPT-4; patients	30 rheumatology questions and physician-generated answers were extracted from the Alberta Rheumatology website; 17 rheumatology patients and 4 rheumatologists rated GPT-4-generated responses and physician-generated responses for comprehensiveness, readability, and overall preference	Patient ratings: GPT-4: mean comprehensiveness (10-point Likert scale): 7.12, readability score: 7.9, physician-generated responses: mean comprehensiveness: 8.76, readability score: 8.75 (no significant differences); rheumatologists' ratings: GPT-4 responses were rated significantly lower for comprehensiveness (5.52 versus 8.76), readability (7.85 versus 8.75), and accuracy (6.48 versus 9.08) compared to physician responses	Although rheumatology patients rated GPT-4-generated responses similarly to physician-generated responses, rheumatologists found GPT-4-generated responses to be inferior, especially in terms of accuracy
74. Mohammad-Rahimi et al., 2023; Germany, Iran, USA, UK[94]	Quantitative	Dentistry; validity and reliability of responses by LLMs on frequently asked questions in the field of endodontics	GPT-3.5, Bard, Bing Chat; patients	20 questions on endodontics were selected, thereof 10 questions developed by 2 endodontists based on their clinical experience and 10 questions provided by GPT-3.5; each prompted 3 times; response evaluation by 2 endodontists using the modified Global Quality Score (GQS) for correctness and content in addition to validity and reliability	Low-threshold validity: GPT-3.5 had the highest validity with 95% valid responses, followed by Bard (85%) and Bing Chat (75%); high-threshold validity: GPT-3.5 had the highest validity with 60% valid responses, Bard and Bing Chat: 15%; Bing Chat had the highest consistency (Cronbach's alpha: 0.955), followed by GPT-3.5 and Bard (Cronbach's alpha: 0.746 and 0.703)	While LLMs show potential as public sources for endodontic information, there are areas for improvement in terms of validity, reliability, and the potential for misinformation
75. Scquizzato et al., 2024; Italy, UK[109]	Quantitative	Emergency Medicine; GPT-3.5's accuracy, relevance, and comprehensiveness in answering questions on cardiac arrest, cardiopulmonary resuscitation (CPR), and post-resuscitation recovery	GPT-3.5; patients	40 questions on different aspects of cardiac arrest were obtained from websites of institutions, scientific societies, and organizations; response evaluation by 8 doctors, 5 nurses, one psychologist, and 16 laypeople for relevance, clarity, comprehensiveness, overall value, and readibility	Overall positive evaluation by professionals and laypeople (5-point Likert scale): 4.3; clarity: 4.4; relevance: 4.3; accuracy: 4; comprehensiveness: 4.2; laypeople rated overall value (4.6 versus 4) and comprehensiveness (4.5 versus 3.9) significantly higher than professionals	GPT-3.5 demonstrates the ability to provide largely accurate, relevant, and comprehensive answers to questions about cardiac arrest, CPR, and post-resuscitation recovery
76. Hermann et al., 2023; USA[95]	Quantitative	Gynecology; GPT-3.5's accuracy in responding to questions on cervical cancer prevention, diagnosis, treatment, and quality of life	GPT-3.5; patients	64 questions adapted from 'frequently asked questions' pages on cancer.net and the American College of Obstetricians and Gynecologists website; response evaluation by 2 attending gynecologic oncologists for correctness and comprehensiveness	Correct and comprehensive: 53.1%; correct but not comprehensive: 29.7%; partially incorrect: 15.6%; completely incorrect: 1.6%	GPT-3.5 accurately answers questions about cervical cancer prevention, survivorship, and quality of life but performs less accurately for questions on cervical cancer diagnosis and treatment

77. Kerbage et al., 2023; USA[96]	Quantitative	Gastroenterology; GPT-4's accuracy in responding to patient questions on irritable bowel syndrome (IBS), inflammatory bowel disease (IBD), colonoscopy, and colorectal cancer (CRC) screening, as well as questions from a physician's perspective on CRC screening and surveillance; to assess GPT-4's ability to generate supportive references for its responses	GPT-4; patients/caregivers	65 questions, thereof 15 questions on colonoscopy and CRC screening, 15 questions on IBS, 20 questions on IBD, and an additional 15 questions on CRC screening and surveillance were designed based on a Google Trends search; a request for references followed each prompt; response grading by 3 gastroenterologists using a granular and an overall grading system for comprehensiveness, accuracy, and references	84% of answers were overall accurate; physician-oriented questions: 47% of the answers were accurate; references were unsuitable for 53% of IBS-related answers, 15% of IBD-related answers, and 27% of colonoscopy and CRC prevention-related answers	GPT-4 shows potential in delivering health information to patients seeking guidance on specific gastrointestinal diseases but should be used with caution for clinical decision-making or as a reference source
78. Shiraishi et al., 2023; Japan[97]	Qualitative	Ophthalmology; GPT-3.5's proficiency in generating accessible informed consent documents for patients undergoing blepharoplasty	GPT-3.5; patients	2 prompts, one short prompt (1) and one detailed prompt with precise instructions (2) were constructed for the creation of consent documents on blepharoplasty; evaluation of GPT-3.5's and original consent documents by 4 board-certified plastic surgeons and 4 nonmedical staff members in terms of accuracy, informativeness, and accessibility	Prompt 1 scored lower than the original consent documents in accuracy (5-point Likert scale: 3.75 versus 5), informativeness (3.75 versus 5), and accessibility (3.25 versus 4.5); prompt 2 scored lower compared to the original IC document in accuracy (4 versus 5) and accessibility (3.25 versus 4.5)	While GPT-3.5 shows potential in generating informed consent documents for blepharoplasty, there are notable differences and limitations compared to the original documents, particularly in terms of accuracy, informativeness, and accessibility
79. Barclay et al., 2023; USA[98]	Quantitative	Ophthalmology; GPT-3.5's and GPT-4's quality and accuracy of information on corneal transplantation and Fuchs dystrophy; to assess whether the answers improve over time	GPT-3.5/GPT-4; patients/caregivers	10 questions on endothelial keratoplasty and Fuchs dystrophy developed by 10 corneal specialists; response evaluation by the same corneal specialists for quality, safety, accuracy, and bias of information on a 1 (A+) to 5 (F) scale	Average score: GPT-4 significantly outscored GPT-3.5 (1.4 versus 2.5); correct facts: GPT-3.5: 61%, GPT-4: 89%, with a significant improvement across iterations; against scientific consensus: GPT-3.5: 35%, GPT-4: 5%	ChatGPT's quality of responses has improved significantly between versions 3.5 and 4, and the likelihood of providing information contrary to scientific consensus has decreased
80. Qarajeh et al., 2023; USA, Jordan, Thailand[99]	Quantitative	Nephrology; effectiveness of different LLMs in accurately determining the potassium and phosphorus content in foods for individuals adhering to a renal diet	GPT-3.5/GPT-4, Bard, Bing Chat; patients	240 dietary items from the Mayo Clinic's renal diet compendium were prompted two times, the second time after a two-week interval, to categorize the items based on their potassium and phosphorus content; response evaluation based on the accuracy of each model	Accuracy in identifying potassium content: GPT-3.5: 66%, GPT-4: 81%, Bard: 79%, Bing: 81%; accuracy rate in identifying phosphorus content: GPT-3.5: 85%, GPT-4: 77%, Bard: 100%, Bing: 89%	LLMs show potential as efficient tools in renal dietary planning, but refinements are warranted for optimal utility
81. Chowdhury et al., 2023; UK[100]	Quantitative	Ophthalmology; to assess the safety and appropriateness of responses generated by GPT-3.5 to post-operative questions from patients who had undergone cataract surgery	GPT-3.5; patients	131 questions collected from automated follow-up calls with a cohort of 120 patients; response evaluation by 2 ophthalmologists using a human evaluation framework adapted from previous work, focusing on helpfulness, clinical harm, and appropriateness	59.9% of responses were "helpful", 36.3% "somewhat helpful", 24.4% had the possibility of "moderate or mild harm", 9.5% were opposed to clinical consensus	GPT-3.5 can potentially address routine patient queries post-cataract surgery safely, but significant safety constraints exist, necessitating careful consideration in healthcare applications
82. Singer et al., 2023; USA[101]	Quantitative	Ophthalmology; to develop and test Aeyeconsult as an ophthalmology chatbot leveraging GPT-4 and verified ophthalmology textbooks to answer eye care-related questions	Aeyeconsult based on GPT-4; patients/caregivers	Aeyeconsult was developed using GPT-4, LangChain, and Pinecone; primary source material was a collection of ophthalmology textbooks in PDF format; 260 ophthalmology questions in multiple-choice format were obtained from OphthoQuestions.com; response comparison against GPT-4 based on 4 categories (correct, incorrect, no answer, multiple answers)	Aeyeconsult outperformed ChatGPT-4 in accuracy (83.4% versus 69.2%) and had fewer no answers (5 versus 18) and multiple answers (0 versus 7)	LLMs can be useful in answering ophthalmologic questions, but their reliability and accuracy are limited due to training on unverified internet data and lack of source citations
83. Xie et al., 2023; Australia[102]	Qualitative	Plastic Surgery; GPT-3.5's utility in simulating doctor-patient consultations for rhinoplasty	GPT-3.5; patients	9 questions based on a rhinoplasty consultation checklist published by the American Society of Plastic Surgeons were posed to GPT-3.5: response evaluation for accuracy, informativeness, and accessibility by 4 plastic surgeons	GPT-3.5 demonstrated an understanding of natural language in a health-specific context, provided coherent, informative, and accessible answers, recognized limitations in providing esoteric and personal advice, was able to assist patients with	GPT-3.5 can be a valuable resource for patients seeking information about rhinoplasty and for surgeons in preoperative assessment and

					basic information about the procedure, its risks, benefits, and outcomes	planning
84. Nastasi et al., 2023; USA[103]	Quantitative	Primary Care; GPT-3.5's appropriateness in responding to questions across various clinical scenarios, including preventive care, acute care, and end-of-life decision-making	GPT-3.5; patients	96 clinical vignettes developed by 4 authors; response evaluation by 2 physicians based on clinical appropriateness, type of recommendation, and consideration of demographic variables	97% of responses were appropriate, 97% appropriately acknowledged uncertainty, and 99% provided appropriate follow-up reasoning; no associations between race or gender with the type of recommendation or with a tailored response; "no insurance" was consistently associated with a specific response related to healthcare costs and access.	GPT-3.5's medical advice was usually safe but often lacked specificity or nuance
85. Sulejmani et al., 2024; USA, Taiwan, France, Germany, Brazil, Poland[111]	Quantitative	Dermatology; GPT-4's ability to provide qualitative and empathetic responses to patient questions about atopic dermatitis (AD) compared to physician responses	GPT-3.5; patients	99 questions were provided by an international group of 11 dermatologists based on commonly asked AD patient questions; response evaluation by the same dermatologists for overall quality and reliability	GPT-3.5 scored an average of 8.18 on a 10-point Likert scale among the raters	GPT-4 may be a valuable resource for providing quality and empathetic responses to patient questions about AD
86. Biswas et al., 2023; Qatar[104]	Quantitative	Discipline not specified; to evaluate the potential of a fine-tuned GPT-3.5-turbo model as a personal medical assistant in the Arabic language	GPT-3.5; patients	Fine-tuning using 5,000 question-answer-pairs on gynecology diseases (4000 training, 1000 validation set) from the Arabic Healthcare Question & Answering dataset; automated performance evaluation using perplexity, coherence, similarity, and token count; human evaluations were conducted by two medical professionals who are native in Arabic, focusing on relevance, accuracy, precision, logic, and originality	Perplexity score: 13.96 (moderate confidence in the model's predictions); average similarity score: 0.1 (low similarity between the generated and original texts); coherence score: 0.33 (moderate coherence in the generated text); mean human evaluation scores (5-point Likert scale): relevance: 3, precision: 3.22, logic: 3.98, originality: 3.94, accuracy: 4.1	GPT-3.5 shows promise in medical assistance applications in Arabic, indicating potential for providing trustworthy medical guidance and enhancing access to healthcare knowledge
87. Panagoulias et al., 2023; Greece[105]	Qualitative	Pulmonology; GPT-4's validity, accuracy, and usefulness in diagnosing tuberculosis based on symptoms described by a human	GPT-4; patients	An evaluation framework was developed; prompt (1) includes a simple symptom description, prompt (2) enquires for more specificity in diagnosing the symptoms, prompt (3) includes more specific or/and diagnostic results if these are requested from the proposed diagnostics suggested by the LLM; the framework was tested on a tuberculosis case	Evaluation answer 1: contextually accurate with correct references; actionable for doctors but generic for patients; economic value was overextended; Evaluation answer 2: contextually generic but with correct references; generic for both doctors and patients; economic value was overextended; Evaluation answer 3: context and references correct; actionable for doctors, precise for patients; exact economic value	GPT-4 performed average to optimum, showing promising results for identifying diseases, assisting doctors and patients, and potentially contributing to the economic cost reductions in the healthcare system
88. Chandra et al., 2023; USA[106]	Quantitative	Dermatology; GPT-3.5's potential to generate allergen-specific patient handouts for allergic contact dermatitis	GPT-3.5; patients	300-word patient handouts about the most common allergies in North America were created by GPT-3.5; evaluation using the Patient Education Materials Assessment Tool for Printable Materials (PEMAT-P) for inaccuracies, erroneous, and misleading information by 2 dermatologists	PEMAT-P understandability score: 79%, actionability score: 60%; factual inaccuracies, erroneous or misleading statements: 2.6	GPT-3.5 may be a useful tool in assisting and generating allergen-specific patient handouts
89. Hung et al., 2023; USA[107]	Quantitative	Plastic Surgery; GPT-3.5's usefulness in generating patient education materials on implant-based breast reconstruction; to compare GPT-3.5-generated with expert-generated materials	GPT-3.5; patients	Patient education materials on implant-based breast reconstruction were generated by 5 breast reconstruction experts and GPT-3.5; evaluation for readability and accuracy by 2 independent reviewers	Expert content had a higher readability (Flesch-Kincaid grade: 7.5 versus 10.5); content accuracy of GPT-3.5: 50%; all incorrect answers were due to information errors	GPT-3.5 can be a powerful tool to generate patient education material, but its readability and accuracy still require improvements

Table 2. Evaluation of included studies according to the Mixed Methods Appraisal Tool (MMAT) 2018.[18]

	Screening questions (for all types)		1. Qualitative					2. Quantitative randomized controlled trials					3. Quantitative nonrandomized				
	S1	**S2**	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**2.1**	**2.2**	**2.3**	**2.4**	**2.5**	**3.1**	**3.2**	**3.3**	**3.4**	**3.5**
1. Samaan et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
2. Eromosele et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
3. Johri et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	N	Y
4. Braga et al.	Y	Y	Y	Y	Y	N	Y	-	-	-	-	-	-	-	-	-	-
5. King et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
6. Huang et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
7. Hanna et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
8. Liu et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
9. Samaan et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	N	Y
10. Patnaik et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
11. Ali et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
12. Suresh et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
13. Yeo et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
14. Knebel et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
15. Zhu et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
16. Lahat et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
17. Bernstein et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
18. Rogasch et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
19. Campbell et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
20. Currie et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
21. Draschl et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
22. Alessandri-Bonetti et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
23. Capelleras et al.	Y	Y	Y	Y	Y	Y	Y	-	-	-	-	-	-	-	-	-	-
24. Coskun et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
25. Durairaj et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
26. Kianian et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
27. Seth et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y

28. Inojosa et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
29. Lyons et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	N	Y
30. Babayiğit et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
31. Mondal et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
32. Kim et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	N	Y
33. Song et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
34. Bitar et al.	Y	Y	-	-	-	-	-	Y	Y	Y	Y	Y	-	-	-	-	-
35. Zalzal et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
36. Chervenak et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
37. Bushuven et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	N	Y	N	Y
38. Jeblick et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
39. Samaan et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
40. Zhou et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
41. Oniani et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
42. Hernandez et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
43. Kuşcu et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
44. Biswas et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
45. Chiesa-Estomba et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
46. Decker et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	Y	Y
47. Kaarre et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
48. Ferreira et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
49. Truhn et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	N	Y
50. Hurley et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
51. Cankurtaran et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
52. Birkun et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	Y	Y
53. Pushpanathan et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
54. Shao et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
55. Vaira et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
56. Chen et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
57. Bellinger et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y

58. Nielsen et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
59. Sezgin et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
60. Floyd et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
61. Uz et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
62. Athavale et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
63. Li et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
64. Seth et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
65. Kuckelman et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
66. Lockie et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
67. Harver et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
68. Li et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
69. Scheschenja et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
70. Gordon et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
71. Stroop et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	C	N	Y
72. Coraci et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	Y	Y
73. Ye et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	Y	Y
74. Mohammad-Rahimi et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
75. Scquizzato et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
76. Hermann et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
77. Kerbage et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
78. Shiraishi et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	Y	N	Y
79. Barclay et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
80. Qarajeh et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
81. Chowdhury et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	N	Y
82. Singer et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
83. Xie et al.	Y	Y	Y	Y	Y	Y	Y	-	-	-	-	-	-	-	-	-	-
84. Nastasi et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
85. Sulejmani et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y
86. Biswas et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	Y	Y
87. Panagoulias et al.	Y	Y	Y	Y	Y	Y	Y	-	-	-	-	-	-	-	-	-	-

88. Chandra et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	N	C	N	Y
89. Hung et al.	Y	Y	-	-	-	-	-	-	-	-	-	-	N	Y	Y	N	Y

Abbreviations: C, Can't tell; N, No; Y, Yes; S1, Are there clear research questions?; S2, Do the collected data allow to address the research questions?; 1.1, Is the qualitative approach appropriate to answer the research question?; 1.2, Are the qualitative data collection methods adequate to address the research question?; 1.3, Are the findings adequately derived from the data?; 1.4, Is the interpretation of results sufficiently substantiated by data?; 1.5, Is there coherence between qualitative data sources, collection, analysis and interpretation?; 2.1, Is randomization appropriately performed?; 2.2, Are the groups comparable at baseline?; 2.3, Are there complete outcome data?; 2.4, Are outcome assessors blinded to the intervention provided?; 2.5, Did the participants adhere to the assigned intervention?; 3.1, Are the participants representative of the target population?; 3.2, Are measurements appropriate regarding both the outcome and intervention (or exposure)?; 3.3, Are there complete outcome data?; 3.4, Are the confounders accounted for in the design and analysis?; 3.5, During the study period, is the intervention administered (or exposure occurred) as intended?
Notes: Categories 4 and 5 are not listed as no studies with quantitative descriptive or mixed methods study designs were identified.

9. 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> (2023).

2 OpenAI. GPT-4 Technical Report. *arXiv preprint* arXiv:2303.08774; 10.48550/arXiv.2303.08774 (2023).

3 Zhao, W. X. *et al.* A survey of large language models. *arXiv preprint* arXiv:2303.18223; 10.48550/arXiv.2303.18223 (2023).

4 Clusmann, J. *et al.* The future landscape of large language models in medicine. *Communications Medicine* **3**, 141; 10.1038/s43856-023-00370-1 (2023).

5 Chen, Z. *et al.* Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint* arXiv:2311.16079; 10.48550/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; 10.48550/arXiv.2402.10373 (2024).

7 Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. *arXiv preprint* arXiv:2402.13178; 10.48550/arXiv.2402.13178 (2024).

8 Yang, X. *et al.* A large language model for electronic health records. *npj Dig Med* **5**, 194; 10.1038/s41746-022-00742-2 (2022).

9 Tian, S. *et al.* Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief Bioinform* **25**; 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; 10.1148/radiol.230725 (2023).

11 McDuff, D. *et al.* Towards accurate differential diagnosis with large language models. *arXiv preprint* arXiv:2312.00164; 10.48550/arXiv.2312.00164 (2023).

12 Jiang, L. Y. *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357-362; 10.1038/s41586-023-06160-y (2023).

13 Liu, S. *et al.* Leveraging Large Language Models for Generating Responses to Patient Messages. *medRxiv* 2023.2007.2014.23292669; 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; 10.1136/bmj.n71 (2021).

16 Ouzzani, M., Hammady, H., Fedorowicz, Z. & Elmagarmid, A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev* **5**, 210; 10.1186/s13643-016-0384-4 (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. *Educ Inf* **34**, 285-291; 10.3233/EFI-180221 (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; 10.1186/s13643-017-0454-2 (2017).

20 Thomas, J. & Harden, A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med Res Methodol* **8**, 45; 10.1186/1471-2288-8-45 (2008).

21 Sociocultural Research Consultants, LLC. *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; 10.2196/49886 (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*; 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**; 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; 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; 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; 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; 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; 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; 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 ChatGPT[1]. *medRxiv* 2023.2008.2028.23294730; 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; 10.1016/j.ajg.2023.08.001 (2023).

33 Patnaik, S. S. & Hoffmann, U. Quantitative evaluation of ChatGPT versus Bard responses to anaesthesia-related queries. *Br J Anaesth* **132**, 169-171; 10.1016/j.bja.2023.09.030 (2024).

34 Ali, H. *et al.* Evaluating the performance of ChatGPT in responding to questions about endoscopic procedures for patients. *iGIE* **2**, 553-559; 10.1016/j.igie.2023.10.001 (2023).

35 Suresh, K. *et al.* Utility of GPT-4 as an Informational Patient Resource in Otolaryngology. *medRxiv* 2023.2005.2014.23289944; 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; 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* **1**, 5-35; 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? *J Transl Med* **21**, 269; 10.1186/s12967-023-04123-5 (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**; 10.3390/diagnostics13111950 (2023).

40 Bernstein, I. A. *et al.* Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. *JAMA Netw Open* **6**, e2330320; 10.1001/jamanetworkopen.2023.30320 (2023).

41 Rogasch, J. M. M. *et al.* ChatGPT: Can You Prepare My Patients for [^{18}F]FDG PET/CT and Explain My Reports? *J Nucl Med*, jnumed.123.266114; 10.2967/jnumed.123.266114 (2023).

42 Campbell, D. J. *et al.* Evaluating ChatGPT Responses on Thyroid Nodules for Patient Education. *Thyroid*®; 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; 10.2967/jnmt.123.266151 (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**; 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. *J Plast Reconstr Aesthet Surg* **87**, 390-402; 10.1016/j.bjps.2023.10.091 (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; 10.1016/j.urology.2023.05.040 (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 Plast Surg Aesthet Med*; 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;* 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; 10.1016/j.jhsa.2023.07.003 (2023).

50 Inojosa, H. *et al.* Can ChatGPT explain it? Use of artificial intelligence in multiple sclerosis communication. *Neurol Res Pract* **5**, 48; 10.1186/s42466-023-00270-8 (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*; 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; 10.7759/cureus.48518 (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; 10.7759/cureus.48296 (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; 10.1016/j.seizure.2023.11.013 (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; 10.1007/s10916-023-02021-3 (2023).

56 Zalzal, H. G., Abraham, A., Cheng, J. H. & Shah, R. K. Can ChatGPT help patients answer their otolaryngology questions? *Laryngoscope Investig Otolaryngol*; 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. *Fertil Steril* **120**, 575-583; 10.1016/j.fertnstert.2023.05.151 (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; 10.1007/s10916-023-02019-x (2023).

59 Jeblick, K. *et al*. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol*; 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; 10.1007/s11695-023-06603-5 (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. *Int J Interact Multimed Artif Intell* **8**, 7-13; 10.9781/ijimai.2023.04.007 (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; 10.7759/cureus.48919 (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; 10.3389/fonc.2023.1256459

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; 10.1111/opo.13207 (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*; 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; 10.1001/jamanetworkopen.2023.36997 (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; 10.1007/s00167-023-07529-2 (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; 10.2196/49280 (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; 10.1038/s41598-023-47500-2 (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*; 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; 10.7759/cureus.46736 (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; 10.1017/s1049023x23006568 (2023).

74 Pushpanathan, K. *et al.* Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. *iScience* **26**, 108163; 10.1016/j.isci.2023.108163 (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; 10.2196/46900 (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*; 10.1002/ohn.489 (2023).

77 Chen, S. *et al.* Use of Artificial Intelligence Chatbots for Cancer Treatment Information. *JAMA Oncol* **9**, 1459-1462; 10.1001/jamaoncol.2023.2954 (2023).

78 Bellinger, J. R. *et al.* BPPV Information on Google Versus AI (ChatGPT). *Otolaryngol Head Neck Surg*; 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; 10.1080/00016489.2023.2254809 (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; 10.2196/49240 (2023).

81 Floyd, W. *et al.* Current Strengths and Weaknesses of ChatGPT as a Resource for Radiation Oncology Patients and Providers. *Int J Radiat Oncol Biol Phys*; 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; 10.1111/1756-185x.14749 (2023).

83 Athavale, A., Baier, J., Ross, E. & Fukaya, E. The potential of chatbots in chronic venous disease patient management. *JVS Vasc Insights* **1**; 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; 10.7759/cureus.40895 (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; 10.1093/asjof/ojad084 (2023).

86 Lockie, E. & Choi, J. Evaluation of a chat GPT generated patient information leaflet about laparoscopic cholecystectomy. *ANZ J Surg*; 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; 10.2214/ajr.23.29622 (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; 10.1016/j.clinimag.2023.06.008 (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. *Cardiovasc Intervent Radiol* **47**, 245-250; 10.1007/s00270-023-03563-2 (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; 10.1016/j.jacr.2023.09.011 (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*; 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**; 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 Rheumatol*; 10.1002/art.42737 (2023).

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; 10.1111/iej.14014 (2024).

95 Hermann, C. E. *et al.* Let's chat about cervical cancer: Assessing the accuracy of ChatGPT responses to cervical cancer questions. *Gynecol Oncol* **179**, 164-168; 10.1016/j.ygyno.2023.11.008 (2023).

96 Kerbage, A. *et al.* Accuracy of ChatGPT in Common Gastrointestinal Diseases: Impact for Patients and Providers. *Clin Gastroenterol Hepatol*; 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*; 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*; 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; 10.3390/clinpract13050104 (2023).

100 Chowdhury, M. *et al.* Can Large Language Models Safely Address Patient Questions Following Cataract Surgery?. *Proceedings of the 5th Clinical Natural Language Processing Workshop*; 10.18653/v1/2023.clinicalnlp-1.17 (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; 10.1016/j.jsurg.2023.11.019 (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; 10.1007/s00266-023-03338-7 (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; 10.1038/s41598-023-45223-y (2023).

104 Biswas, M., Islam, A., Shah, Z., Zaghouani, W. & Brahim Belhaouari, S. Can ChatGPT be Your Personal Medical Assistant?. *Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 1-5; 10.1109/SNAMS60348.2023.10375477 (2023).

105 Panagoulias, D., Palamidas, F., Virvou, M. & Tsihrintzis, G. Evaluating the Potential of LLMs and ChatGPT on Medical Diagnosis and Treatment. *14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 1-9; 10.1109/IISA59645.2023.10345968 (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; 10.1089/derm.2023.0059 (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. *Ann Plast Surg* **91**, 409-412; 10.1097/SAP.0000000000003634 (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*; 10.1055/a-2219-4901 (2024).

109 Scquizzato, T. *et al*. Testing ChatGPT ability to answer laypeople questions about cardiac arrest and cardiopulmonary resuscitation. *Resuscitation* **194**, 110077; 10.1016/j.resuscitation.2023.110077 (2024).

110 Kuckelman, I. J. *et al*. Assessing AI-Powered Patient Education: A Case Study in Radiology. *Acad Radiol* **31**, 338-342; 10.1016/j.acra.2023.08.020 (2024).

111 Sulejmani, P. *et al*. A large language model artificial intelligence for patient queries in atopic dermatitis. *J Eur Acad Dermatol Venereol*; 10.1111/jdv.19737 (2024).

112 Currie, G. & Barry, K. ChatGPT in Nuclear Medicine Education. *J Nucl Med Technol* **51**, 247-254; 10.2967/jnmt.123.265844 (2023).

113 Currie, G. M. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy? *Semin Nucl Med* **53**, 719-730; 10.1053/j.semnuclmed.2023.04.008 (2023).

114 Li, J., Dada, A., Puladi, B., Kleesiek, J. & Egger, J. ChatGPT in healthcare: A taxonomy and systematic review. *Comput Methods Programs Biomed* **245**, 108013; 10.1016/j.cmpb.2024.108013 (2024).

115 Jin, M. *et al*. Health-LLM: Personalized Retrieval-Augmented Disease Prediction Model. *arXiv preprint arXiv:2402.00746*; 10.48550/arXiv.2402.00746 (2024).

116 Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint* arXiv:2303.13375; 10.48550/arXiv.2303.13375 (2023).

117 Brin, D. *et al*. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep* **13**, 16492; 10.1038/s41598-023-43436-9 (2023).

118 Jung, L. B. *et al*. ChatGPT Passes German State Examination in Medicine With Picture Questions Omitted. *Dtsch Arztebl Int* **120**, 373-374; 10.3238/arztebl.m2023.0113 (2023).

119 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; 10.1148/radiol.230582 (2023).

120 Singhal, K. *et al*. Towards expert-level medical question answering with large language models. *arXiv preprint* arXiv:2305.09617; 10.48550/arXiv.2305.09617 (2023).

121 Kung, T. H. *et al*. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* **2**, e0000198; 10.1371/journal.pdig.0000198 (2023).

122 Kapoor, S., Henderson, P. & Narayanan, A. Promises and pitfalls of artificial intelligence for legal applications. *arXiv preprint* arXiv:2402.01656; 10.48550/arXiv.2402.01656 (2024).

123 Navigli, R., Conia, S. & Ross, B. Biases in Large Language Models: Origins, Inventory, and Discussion. *ACM J Data Inf Qual* **15**, 1-21; 10.1145/3597307 (2023).

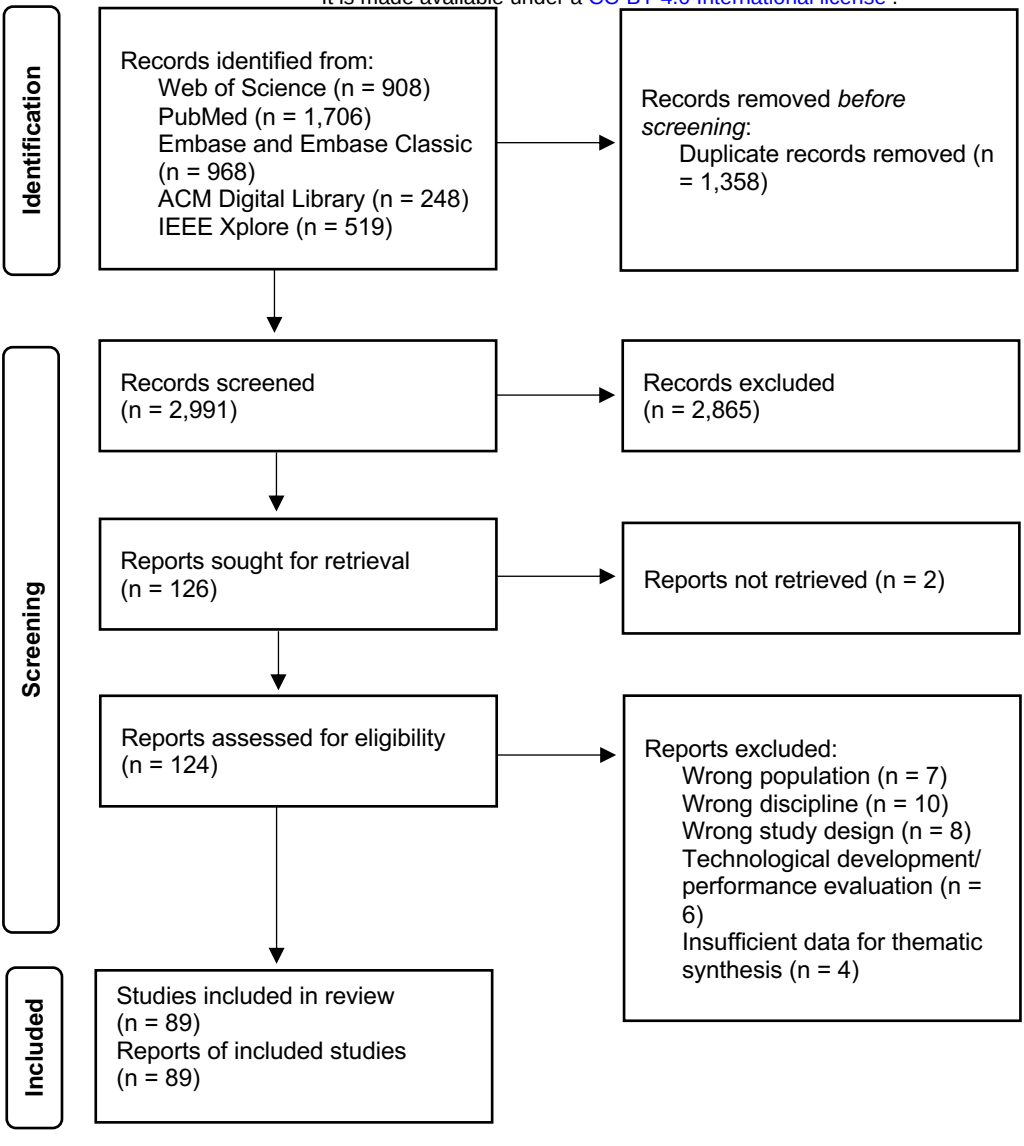
124 Deng, G. *et al*. Jailbreaker: Automated jailbreak across multiple large language model chatbots. *arXiv preprint* arXiv:2307.08715; 10.48550/arXiv.2307.08715 (2023).

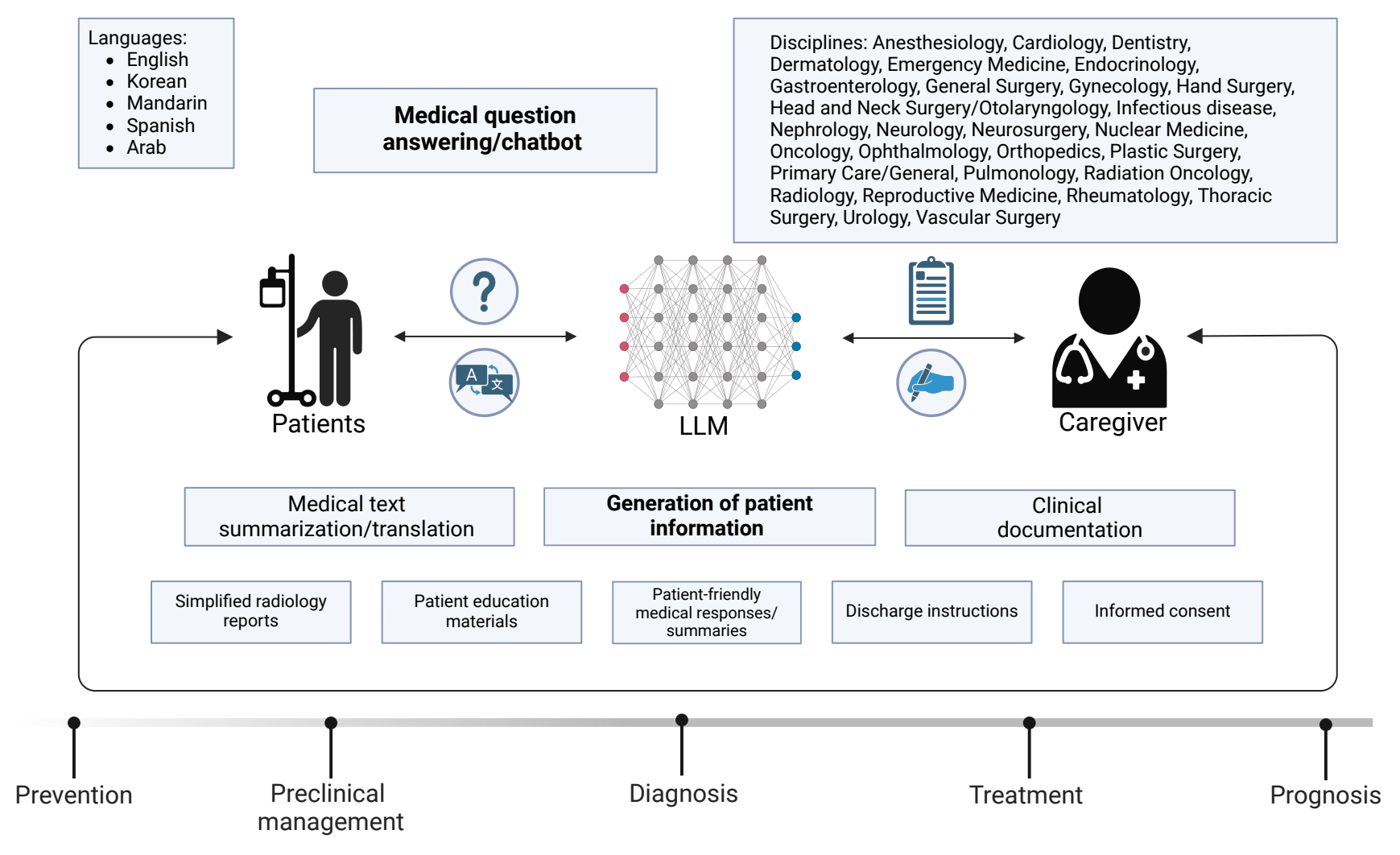
125 Ayoub, N. F., Lee, Y. J., Grimm, D. & Divi, V. Head-to-Head Comparison of ChatGPT Versus Google Search for Medical Knowledge Acquisition. *Otolaryngol Head Neck Sur*g; 10.1002/ohn.465 (2023).

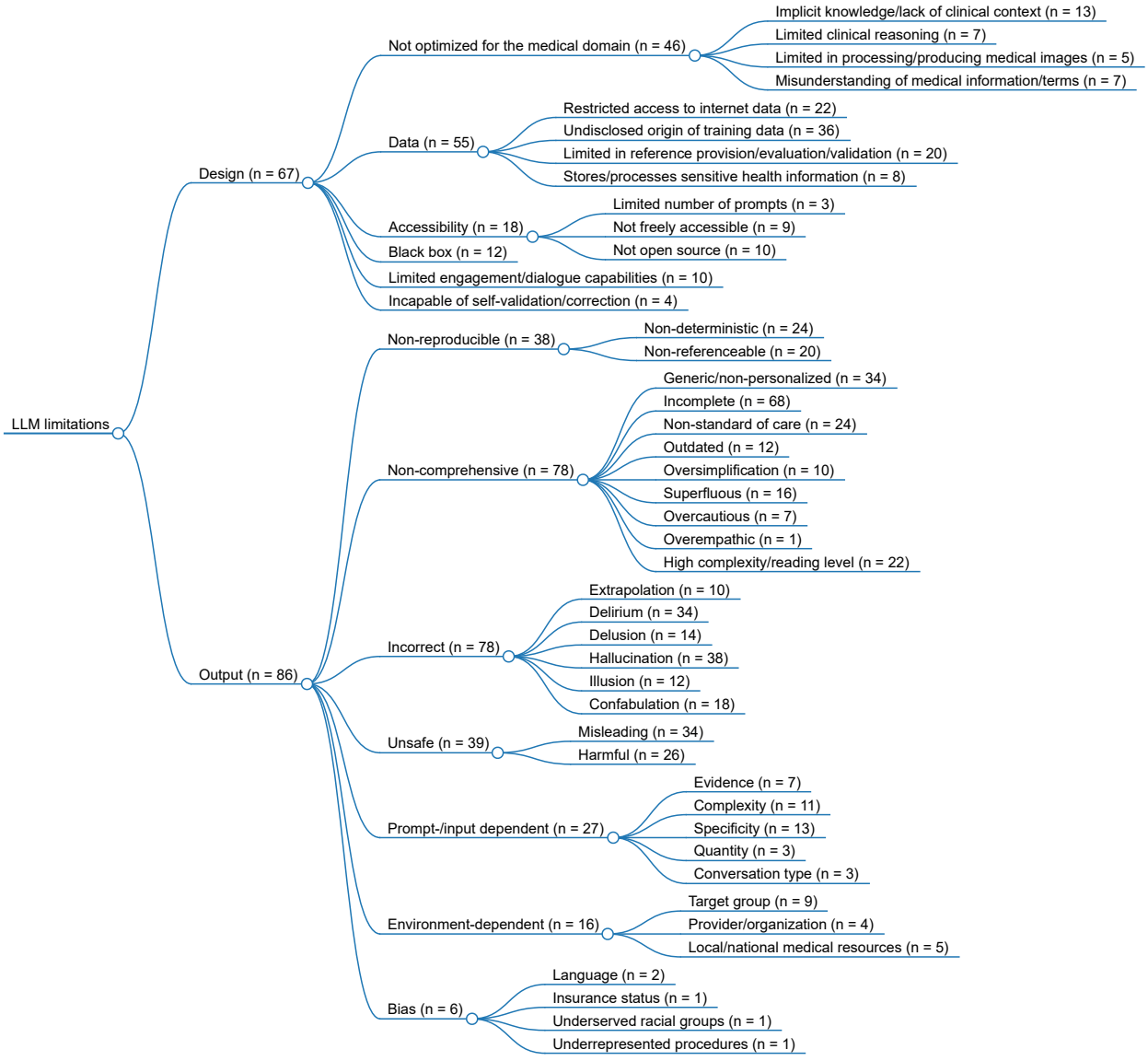
126 Weis, B. Health Literacy: A Manual for Clinicians. Chicago, IL: American Medical Association, American Medical Foundation (2003).

127 Tierney, A. A. *et al.* Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. *NEJM Catalyst* **5**, CAT.23.0404; 10.1056/CAT.23.0404 (2024).

128 Council of the European Union. *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*. 2021/0106(COD), 1-272 (2024).




Languages:
English
Korean
Mandarin
Spanish
Arab
Medical question answering/chatbot
Disciplines: Anesthesiology, Cardiology, Dentistry, Dermatology, Emergency Medicine, Endocrinology, Gastroenterology, General Surgery, Gynecology, Hand Surgery, Head and Neck Surgery/Otolaryngology, Infectious disease, Nephrology, Neurology, Neurosurgery, Nuclear Medicine, Oncology, Ophthalmology, Orthopedics, Plastic Surgery, Primary Care/General, Pulmonology, Radiation Oncology, Radiology, Reproductive Medicine, Rheumatology, Thoracic Surgery, Urology, Vascular Surgery
Patients
LLM
Caregiver
Medical text summarization/translation
Generation of patient information
Clinical documentation
Simplified radiology reports
Patient education materials
Patient-friendly medical responses/ summaries
Discharge instructions
Informed consent
Prevention
Preclinical management
Diagnosis
Treatment
Prognosis


LLM limitations
Design (n = 67)
Not optimized for the medical domain (n = 46)
Implicit knowledge/lack of clinical context (n = 13)
Limited clinical reasoning (n = 7)
Limited in processing/producing medical images (n = 5)
Misunderstanding of medical information/terms (n = 7)
Data (n = 55)
Restricted access to internet data (n = 22)
Undisclosed origin of training data (n = 36)
Limited in reference provision/evaluation/validation (n = 20)
Stores/processes sensitive health information (n = 8)
Accessibility (n = 18)
Limited number of prompts (n = 3)
Not freely accessible (n = 9)
Not open source (n = 10)
Black box (n = 12)
Limited engagement/dialogue capabilities (n = 10)
Incapable of self-validation/correction (n = 4)
Output (n = 86)
Non-reproducible (n = 38)
Non-deterministic (n = 24)
Non-referenceable (n = 20)
Non-comprehensive (n = 78)
Generic/non-personalized (n = 34)
Incomplete (n = 68)
Non-standard of care (n = 24)
Outdated (n = 12)
Oversimplification (n = 10)
Superfluous (n = 16)
Overcautious (n = 7)
Overempathic (n = 1)
High complexity/reading level (n = 22)
Incorrect (n = 78)
Extrapolation (n = 10)
Delirium (n = 34)
Delusion (n = 14)
Hallucination (n = 38)
Illusion (n = 12)
Confabulation (n = 18)
Unsafe (n = 39)
Misleading (n = 34)
Harmful (n = 26)
Prompt-/input dependent (n = 27)
Evidence (n = 7)
Complexity (n = 11)
Specificity (n = 13)
Quantity (n = 3)
Conversation type (n = 3)
Environment-dependent (n = 16)
Target group (n = 9)
Provider/organization (n = 4)
Local/national medical resources (n = 5)
Bias (n = 6)
Language (n = 2)
Insurance status (n = 1)
Underserved racial groups (n = 1)
Underrepresented procedures (n = 1)