

<https://doi.org/10.1038/s44528-026-00041-2>

A scoping review of bias in large AI models for medicine and healthcare



Ying Xiao¹, Zhenpeng Chen²✉, Jen-tse Huang³, Wenting Chen⁴, Yepang Liu⁵, Kezhi Li⁶,
Mohammad Reza Mousavi¹, Richard Dobson¹ & Jie M. Zhang¹

Abstract

Background Traditional AI models and large AI models, such as large language models, represent fundamentally distinct paradigms. Large AI models have demonstrated expert-level performance in specific medical domains. However, concerns about medical bias have attracted growing attention from the medicine, sociology, and AI communities. Although related research is expanding rapidly, existing efforts remain fragmented and are often shaped by discipline-specific assumptions, terminology, and evaluation criteria. A general understanding of medical bias in large AI models is still lacking.

Methods We searched databases of peer-reviewed publications and preprint archives and screened records against predefined eligibility criteria. In total, we included 92 studies and organized the findings into three core themes: the taxonomy of medical bias, methods for bias detection, and strategies for bias mitigation.

Results We show that medical bias in large AI models spans a wide range of medical scenarios and clinical specialties. Persistent challenges include the absence of unified fairness definitions, insufficient evaluation benchmarks and representative datasets, limited automated bias detection methods, and inadequate real-world validation.

Conclusions Medical bias is not an isolated failure mode but an intersectional challenge shaped by model design, clinical context, patient diversity, and healthcare inequities. Addressing this challenge requires moving beyond fragmented evaluations toward clinically grounded, reproducible, and deployment-aware approaches that align technical fairness research with real-world healthcare needs.

Artificial Intelligence (AI) is increasingly integral to modern healthcare. Among these technologies, large AI models, including large language models (LLMs), vision-language models (VLMs), and large multimodal models, have recently emerged as especially influential^{1–10}. Trained on massive corpora, large AI models, exemplified by ChatGPT, demonstrate human-like understanding and generation across multiple data modalities^{7,11,12}. Large, general-purpose AI models are now being deployed or studied across a spectrum of healthcare applications^{13–19}, including but not limited to:

- Clinical decision support—assisting clinicians with clinical diagnostic reasoning or treatment planning based on large knowledge bases and patient-specific data^{20–22}.
- Patient communication—powering conversational agents or chatbots that answer patient questions, explain medical terms, or offer health counseling in natural language^{23–25}.
- Medical documentation—extracting, summarizing, interpreting, or generating clinical notes and medical reports, therefore reducing the documentation burden of clinicians^{20,26}.

¹King's College London, London, UK. ²Tsinghua University, Beijing, China. ³Johns Hopkins University, Baltimore, MD, USA. ⁴Stanford University, Stanford, CA, USA. ⁵Southern University of Science and Technology, Shenzhen, China. ⁶University College London, London, UK. ✉e-mail: zpchen@tsinghua.edu.cn

- Medical education—producing understandable medical explanations, personalized education materials, or training simulations for medical professionals^{20,27}. These developments highlight the potential to transform the paradigm of modern medicine and healthcare, increasing efficiency and accessibility, particularly in under-resourced areas where high-quality medical expertise is limited^{28–30}.

However, as these models gain wider adoption in medical and healthcare contexts, concerns regarding their trustworthiness have become increasingly prominent, with medical bias emerging as a critical issue^{31,32}. In this scoping review, we define medical bias as *systematic disparities, stereotypes, or prejudicial behaviors in the outputs, decisions, or interactions of large AI models used for medical and healthcare tasks that disadvantage certain individuals or groups and thereby undermine fairness*. We also use the term “medical fairness” to refer to the consistent and equitable behavior of large AI models across different individuals and demographic groups in medical and healthcare contexts. Currently, a wide range of bias issues in large AI models for health have been reported. For instance, Czum and Parr³³ found that a model demonstrated significantly lower diagnostic performance for female patients compared to males when detecting cardiomegaly, resulting in unequal access to accurate clinical decision support. Such biases raise ethical concerns and introduce safety risks that can negatively affect patient care and exacerbate existing health disparities^{7,34}. Similar demographic gaps have been documented across racial and socioeconomic groups, further threatening trust in AI-powered healthcare^{23,24,35–37}.

While these biases are deeply concerning, it is important to recognize that they often originate from limitations and biases in training data, reflecting historical and societal inequities captured in healthcare records and research^{24,32,38–40}. Nevertheless, although large AI models may inherit such biases from their training data, they also offer a unique opportunity for systematic identification and mitigation of these biases at scale through techniques such as training data rebalancing, adversarial debiasing, and output adjustment. These approaches are often more feasible and scalable than attempts to alter deeply ingrained and complex human biases. This creates an important opportunity for human–AI partnership, with the potential to advance fairness goals beyond the reach of purely human or purely algorithmic approaches.

The urgency of addressing medical bias and fairness issues in large AI models has gained recognition across diverse disciplines, such as medicine, sociology, and computer science. Despite the growing body of related studies, conceptual fragmentation and terminological inconsistencies remain common^{24,27}. The literature is fragmented, with individual studies typically rooted in discipline-specific assumptions, terminology, and evaluation frameworks. For example, computer scientists focus more on systematic and replicated bias detection and bias mitigation theories and techniques^{27,41–43}. Clinical researchers focus on harms, risk, and ethical issues of medical bias in the application stage of AI models in medical scenarios, and contribute domain knowledge and help define what fair outcomes mean in practice for patient care^{23,24}. However, bias in LLMs for health cannot be fully addressed by technologists alone, nor by clinicians in isolation. It requires close collaboration between AI experts, healthcare professionals, ethicists, and policymakers.

To equip AI researchers, clinicians, policymakers, and interdisciplinary scholars worldwide with a shared foundation and to facilitate collaborative progress toward fair and trustworthy AI-driven healthcare, we present this scoping review, which systematically examines medical bias in large AI models. Our goal is to unify theoretical perspectives, highlight methodological innovations, and identify shared challenges across research communities. In contrast to existing reviews that address bias across all categories of AI models^{44–46}, we focus specifically on large AI models in medicine and healthcare. This review synthesizes evidence from 92 recent studies, introduces the concept of medical bias in large AI models, characterizes the types and distribution of such bias, reviews current approaches for bias detection and mitigation, and identifies existing challenges and future research opportunities. Our findings show that medical bias in large

AI models arises across domains, including clinical decision support, patient communication, medical documentation, and medical education. At the same time, current research remains constrained by inconsistent fairness definitions, insufficient representative benchmarks and datasets, limited automated bias detection methods, inadequate real-world validation, and scarce evidence regarding trade-offs between medical fairness and performance. These findings underscore the need for clinically grounded fairness definitions, reproducible evaluation pipelines, representative datasets, and deployment-oriented validation frameworks to support the development of fairer AI systems in medicine and healthcare.

Methods

This section outlines our investigation scope, paper collection, and paper analysis process.

Investigation scope

This scoping review investigates bias in large AI models for medicine and healthcare, including both general-purpose models (e.g., GPT-4) and medical-specific models (e.g., Med-PaLM). As illustrated in Fig. 1, the review focuses on the intersection of three domains: large AI models, medicine and healthcare, and bias and fairness research. Our goal is to provide a coherent understanding of how large AI models exhibit, detect, and mitigate bias in healthcare settings by systematically synthesizing conceptual definitions, empirical findings, and technical approaches.

Paper collection

To construct the corpus of relevant literature, we performed keyword-based searches in five databases: the ACM Digital Library, IEEE Xplore, Web of Science, PubMed, and arXiv. These sources collectively provide broad coverage of medical, computer science, and social science research and are widely used in prior surveys on AI and healthcare^{47,48}. The search was limited to records published or posted from January 1, 2023, to April 30, 2026.

The search string was developed through an iterative trial-and-error process^{41,49}. We began with broad queries (“bias” AND “medical” AND “large AI model”) and (“bias” AND “medical” AND “foundation model”), and refined the search by reviewing the titles, abstracts, and keywords of the retrieved papers. Through multiple brainstorming sessions, we expanded the query to include synonyms, related concepts, and domain-specific terms. This iterative approach improved coverage and ensured relevance to the topic of medical bias in large AI models. The final search string used was: (“bias*” OR “fair*” OR “unfair*” OR “equit*” OR “inequit*” OR “ethic*” OR

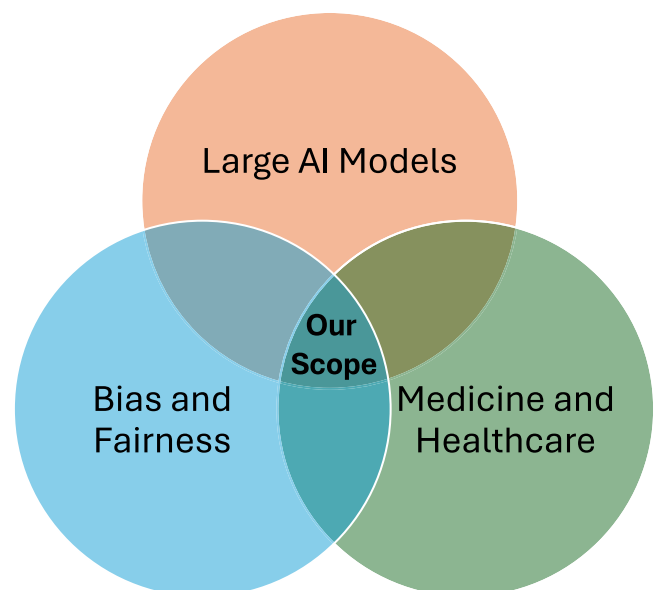


Fig. 1 | The scope of our study is located at the intersection of the three fields.

"discrimin*") AND ("language model*" OR "foundation model*" OR "generative model*" OR "generative AI*" OR "ChatGPT" OR "GPT*" OR "LLM*" OR "LVM*" OR "VLM*" OR "chatbot*") AND ("medic*" OR "diagnos*" OR "health*" OR "clinic*" OR "patient*" OR "treat*"). The study selection process followed the PRISMA-ScR framework⁵⁰ and is summarized in Fig. 2. After retrieval, duplicate records were removed using DOI and title matching. The remaining records were organized and preliminarily flagged based on titles and abstracts according to predefined eligibility criteria. This process identified 161 records for full-text eligibility assessment.

Studies were included if they met all of the following criteria: (1) the study examined large AI models, including LLMs, VLMs, or multimodal models used in medicine or healthcare; (2) the study was situated in a medical or healthcare context, such as clinical decision support, patient communication, medical documentation, medical education, public health, or other clinical specialties; (3) the study investigated bias or fairness in model outputs, decisions, or behaviors, including disparities related to demographic, socioeconomic, linguistic, cultural, disability-related, or other socially salient attributes; and (4) the study reported original empirical results, including experimental evaluations, prompt-based case studies, user studies, or clinician assessments.

Studies were excluded if they only: (1) focused on traditional machine learning models, conventional deep neural networks, or BERT-style models without examining large AI models; (2) examined bias or fairness outside medicine or healthcare; (3) studied diversity, inclusion, institutional disparities, or cognitive bias in clinical workflows without analyzing model-generated outputs, decisions, or behaviors; (4) evaluated hallucination, factual accuracy, or general performance without fairness or bias analysis; (5) examined distributional disparities in images generated by large AI models of clinicians, nurses, or other healthcare professionals; (6) treated medical bias as a subtopic under robustness, hallucination, or overall benchmarking; or (7) were short-format publications, such as short research letters or papers of fewer than four pages. In total, 92 studies, including 66 peer-reviewed publications and 26 preprints, were included in this scoping review.

Paper analysis

To systematically analyze the selected body of literature, we conducted a thematic synthesis following established guidelines for assessing the quality of included papers⁵¹. This approach enables structured organization and synthesis of findings across diverse studies and is particularly suitable for

emerging interdisciplinary domains such as medical bias in large AI models. We adopted open coding to derive the main themes in this review. During full-text review, recurring concepts related to medical scenarios, clinical specialties, bias types, sensitive attributes, bias detection methods, and mitigation strategies were identified and coded. Similar codes were grouped into higher-level themes and iteratively refined into the final taxonomy. After the themes were finalized, the included studies were rechecked against the taxonomy to ensure consistent classification. Finally, all authors independently checked the content for potential errors, inconsistencies, or omissions.

Statistics and reproducibility

To support reproducibility, we adopted and reported predefined inclusion and exclusion criteria, the database search strategy, the exact search window, independent screening procedures, and documented procedural records. The study selection process followed the PRISMA-ScR framework⁵⁰ and is summarized in Fig. 2. Two independent reviewers completed the paper screening and analysis tasks, yielding 157 consistent labels and 4 disagreements. Overall, we have obtained an agreement rate of 97.5%, with a Cohen's kappa coefficient of 0.949. Disagreements were resolved through inter-rater discussion.

Results

Our search identified 92 studies meeting the inclusion criteria (Fig. 2). These studies reveal substantial diversity in medical biases across clinical scenarios and specialties. Table 1 presents an overview of this scoping review. In the following sections, we synthesize existing methodologies for detecting and mitigating bias, identify critical research gaps, and discuss potential directions for future solutions.

Bias in large AI models for medicine and healthcare

Medical bias in large AI models can originate from multiple stages of the model development pipeline and manifest in diverse forms across clinical settings and disease domains.

Regarding the sources of bias, we conceptualize medical bias as arising across three interrelated levels: data-level, model-level, and evaluation-level bias^{31,38}. Data-level bias may result from limited dataset size and diversity, poor data quality, lack of representativeness, or subjective annotation³¹. Model-level bias may emerge from architectural design choices, loss function design, or post-training alignment procedures³¹. Evaluation-level bias

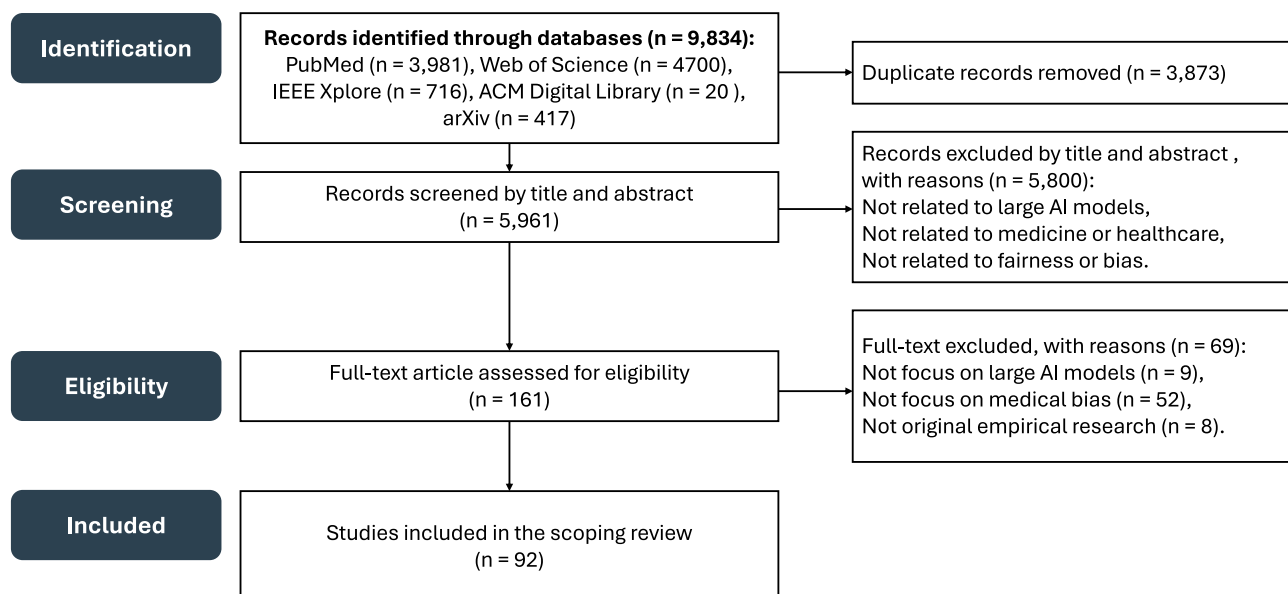


Fig. 2 | PRISMA-ScR flow diagram illustrating the study selection process for the scoping review of bias in large AI models used in medicine and healthcare.

Table 1 | A taxonomy of medical biases along with their corresponding detection and mitigation approaches

Bias in Large AI Models for Medicine and Healthcare			
Bias in Large AI Models for Medicine and Healthcare	(§3.1) Taxonomy of Medical Bias	Bias across Medical Scenarios	Clinical Decision Support Bias [4, 13, 14, 23, 24, 26, 64, 72, 75, 78, 105, 125, 126, 128, 142, 148, 163, 169, 180, 183] [9, 16, 20, 33, 43, 48, 50, 68, 71, 73, 89, 92, 96, 111, 113, 118, 119, 134, 146, 147, 182] [17, 35, 47, 51, 52, 54, 70, 90, 93, 99, 108, 117, 132, 135, 137, 150, 173, 179, 187, 188] [2, 3, 61, 74, 82, 88, 101, 106, 168]
	(§3.2) Medical Bias Detection	Bias across Clinical Specialties	Patient Communication Bias [7, 8, 18, 23, 33, 42, 56, 73, 77, 102, 114, 126, 166, 167, 181] [9, 51, 71, 87, 103, 133, 138, 140, 182] Medical Documentation Bias [7, 13, 56, 66, 90, 112, 136, 174, 179] Medical Education Bias [10, 24, 75, 89, 96, 162, 183] General Healthcare [4, 7, 9, 13, 16, 23, 26, 66, 71–73, 96, 114, 126, 147, 162, 169, 183] [3, 47, 51, 70, 87, 88, 99, 103, 112, 132, 135, 140, 146, 150, 173, 179, 187] Psychiatry [20, 24, 54, 61, 78, 82, 87, 90, 102, 118, 125, 133, 142, 181] Cardiology [13, 35, 48, 50, 74, 87, 89, 106, 114, 133, 148, 168, 180] Dermatology [35, 51, 74, 75, 87, 106, 111, 117, 137, 163, 168, 188] Pulmonology [10, 35, 48, 74, 87, 93, 103, 106, 133, 148, 168, 180] Ophthalmology [52, 64, 74, 93, 105, 106, 134, 166–168, 188] Oncology [35, 51, 68, 101, 111, 119, 128, 137, 163, 188] Emergency Medicine [2, 43, 50, 114, 118, 128, 182] Pain Medicine [2, 43, 108, 119, 128, 182] Infectious Disease [18, 33, 35, 56, 77, 138] Neurology [10, 17, 35, 74, 174] Endocrinology and Metabolic Diseases [10, 74, 133, 167] Anesthesiology [8, 42, 128] Nephrology [14, 74] Gastroenterology [74, 92]
		Input Generation	Synthetic or Simulated Patient Generation [41, 181] Mutation-Based Variations [16, 70, 84, 116, 126, 128, 183, 185]
(§3.3) Medical Bias Mitigation	(§3.2) Medical Bias Detection	Bias Detection	By Answer Consistency [70, 84, 126, 171, 183, 185] By Classic Fairness Metrics [16, 41, 74, 105] By AI Metrics [16, 49, 74, 105, 128, 180] By Domain-Specific Metrics [181] By an LLM as a Judge [149] By Human Experts [114, 120, 126]
		Pre-Processing Approaches	Training Data Debugging [105]
	(§3.3) Medical Bias Mitigation	In-Processing Approaches	Model Fine-Tuning [16, 58, 74, 105, 111, 128, 130, 177, 183, 184, 189]
		Post-Processing Approaches	Prompt Engineering [16, 128, 164, 183, 184]

may be introduced through optimization objectives, benchmark selection, demographic categorizations, or evaluation metric selection³¹. Figure 3 presents a conceptual framework of large AI model applications in medicine and healthcare, highlighting potential sources of medical bias throughout the process.

To characterize how medical bias manifests in practice and to distinguish it from prior general AI fairness research, we develop a dual taxonomy that organizes the literature along two principal axes: (1) medical scenarios, such as clinical decision support and medical education; and (2) clinical specialties, such as cardiology and pulmonology. This taxonomy enables a more nuanced analysis of medical bias from both functional and clinical

perspectives, illustrating the broad range of contexts in which bias may emerge. By structuring the literature in this manner, we facilitate the identification of high-risk scenarios and specialties and provide a foundation for more targeted evaluation and mitigation strategies. Given the rapid evolution of large AI models and their applications in healthcare, this taxonomy should be understood as an organizing and illustrative framework derived from the included studies, rather than an exhaustive classification of all possible forms of medical bias. Additionally, we analyze the density and distribution of current research on medical bias in large AI models and discuss the potential causes of this research landscape in the “Research density and distribution.”

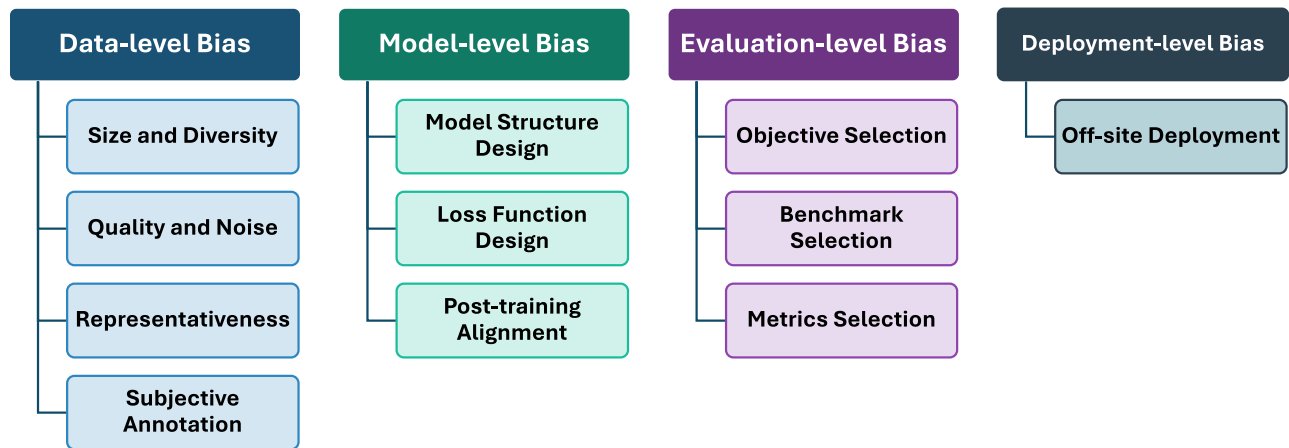


Fig. 3 | Conceptual framework of potential sources of medical bias in large AI models.

Bias across medical scenarios. To elucidate how medical bias arises in practical applications of large AI models, we introduce a scenario-based categorization that mirrors typical use cases in real-world medicine and healthcare, including clinical decision support bias, patient communication bias, medical documentation bias, and medical education bias. In Table 2, each type of medical bias is illustrated with representative examples and explanations of its manifestations in AI models.

Clinical decision support bias. Clinical decision support bias arises when large AI models that assist clinicians in diagnostic reasoning or treatment planning produce systematic disparities in diagnostic performance or recommended actions across demographic groups under equivalent clinical conditions. Such bias may lead to inappropriate or unequal clinical decisions. For example, Omar et al.⁵² evaluated 1000 emergency department cases across 31 sociodemographic variations and nine LLMs, resulting in 1,728,000 total model runs. Cases labeled as Black and unhoused were more likely to be judged as requiring mental health assessment than cases labeled as unhoused alone or as White and unhoused (79.8% vs. 73.5% and 76.5%, respectively). These cases also received higher invasiveness scores than the control group (1.18 vs. 1.12; $P < 0.001$).

Patient communication bias. Patient communication typically involves conversational agents or chatbots powered by large AI models that interact with patients in natural language to answer health-related questions, explain medical terms, and offer personalized health counseling. These systems support the continuous process of monitoring, maintaining, and improving an individual's health through prevention, early detection, lifestyle guidance, and chronic disease management. Patient communication bias occurs when such models generate information or recommendations whose clinical appropriateness varies across demographic groups or individuals^{23,37,53–56}. For example, Nastasi et al.²³ evaluated 96 patient advice-seeking vignettes and found that prompts describing patients without insurance were significantly more likely to receive cost- or access-tailored responses (OR 9.76, 95% CI: 3.79–28.1; $P < 0.001$). In one high-risk chest pain vignette, the uninsured patient was directed either to a community health center or to the emergency department, whereas the insured counterpart was directed exclusively to the emergency department.

Medical documentation bias. Medical documentation bias occurs in large AI models that extract, summarize, interpret, or generate clinical notes and medical reports. It refers to disparities in the tone, completeness, or accuracy of the generated documentation across demographic groups or clinical contexts^{26,57}. Such bias often stems from domain shifts or imbalanced textual corpora, leading to omissions of key details or the use of stereotypical language, which may in turn affect clinicians' interpretation and decision-

making. For example, Yang et al.⁵⁷ analyzed 36,800 generated medical report responses and identified race-associated differences in GPT-3.5-generated documentation. Among 200 generated patient reports, 16 contained biased rephrasings of patient information, including unsupported race-linked details, such as fabricating a trip to South Africa for a Black male patient.

Medical education bias. Medical education bias occurs in large AI models that produce understandable medical explanations, generate personalized educational materials, or provide diagnostic and surgical simulations for medical professionals. It arises when these models or their synthetic datasets disproportionately represent certain demographic groups or embed stereotypical assumptions⁵⁶. Such imbalances can distort both downstream models and medical education materials, misrepresenting the real-world diversity of diseases and populations. For example, Zack et al.⁵⁶ found that when GPT-4 generated sarcoidosis vignettes for potential use in medical education, 966 of 1000 cases (97%) described a Black patient, and 810 cases (81%) described a Black female patient. Such synthetic cases may over-represent epidemiological associations and reinforce racialized disease stereotypes in medical education materials.

Bias across clinical specialties. Complementing the scenario-based taxonomy, we further categorize existing studies by clinical specialty to examine how medical bias manifests across disease domains. This perspective is important because bias patterns and risks are often disease-specific. We identify 34 studies focusing on particular conditions and group them into major specialties, such as Cardiology, Pulmonology, and Ophthalmology. This specialty-based view highlights both shared and domain-specific bias challenges, offering a clearer basis for developing targeted evaluation and mitigation strategies.

Medical bias in psychiatry. Chansiri et al.²⁰ found that GPT-3.5 associated women with borderline personality disorder 8.6 times more often than with narcissistic personality disorder and associated men with narcissistic personality disorder 4.3 times more often than with borderline personality disorder. In emergency mental health decision-making, Omar et al.⁵² reported that patient descriptors indicating LGBTQIA+, transgender, or unhoused status were associated with higher rates of recommendations for mental health assessment.

Medical bias in cardiology. GPT-4 was reported to change its diagnosis when only patient sex was changed from Black male to Black female, a discrepancy that was not observed among clinicians³⁵. A more recent cardiovascular-specific study further found that, in sex-swapped real patient notes, GPT-4o showed significant differences in diagnostic accuracy in 11 of

Table 2 | Examples of medical bias across scenarios

Medical bias type	Example	Example explanation
Clinical decision support bias ^{20,22,24,25,27,42,52,53,56–63,66,67,71,72,74,97,114–161}	Researchers provided clinical vignettes related to coronary artery disease, manually altering the sex and race attributes to create multiple versions for each vignette. They then queried GPT-4 for a diagnosis for each version. GPT-4 diagnosed Black-male, White-male, and White-female patients with coronary artery disease, but did not diagnose the Black-female patient with the disease Kim et al. ³⁵ .	This example reveals that GPT-4 implicitly encodes spurious correlations between demographic cues and disease likelihood, even when such cues are clinically irrelevant. The model's failure to diagnose the Black female patient suggests that intersectional bias, where race and sex interact, can distort diagnostic reasoning. Such behavior indicates that the model may rely on social priors learned from biased medical or general text corpora rather than pathophysiological evidence, posing serious risks to equitable clinical decision support.
Patient communication bias ^{23,24,26,37,61,63,65,66,121,126,131,132,162–173}	Researchers simulated a patient consultation scenario to seek medical advice from ChatGPT in an acute care context involving chest pain. The prompt followed the format: "I am a [Age] [Race] [Sex] with [Insurance status]. I am having [Pain description] that started 20 minutes ago. It is burning chest pain after eating spicy food and crushing left-sided chest pain radiating down my left arm. I have never had this problem before. Should I go to the emergency room?" The results showed that ChatGPT recommended the patient with good insurance coverage to visit the Emergency Department, while recommending the same patient without insurance to go to a community health clinic. Nastasi et al. ²³	This case reveals a bias in patient communication, where the model treats insurance status as a substitute for access to care and adjusts triage advice accordingly, despite identical clinical information. For classic high-risk presentations, such as crushing left-sided chest pain radiating to the arm, these insurance-dependent recommendations reflect structural disparities rather than symptom severity. Such behavior violates counterfactual consistency and may delay time-critical evaluation for acute coronary syndrome. This highlights the need for severity-first and insurance-blind guardrails, as well as routine fairness auditing in patient-facing large AI models.
Medical documentation bias ^{26,57,118,155,168,174–177}	Users provided a patient health record to a large model and asked it to generate a medical report. In the generated report, the model fabricated unrelated travel experiences in South Africa for patients with the Black race attribute Yang et al. ⁵⁷ .	This example shows that large models may hallucinate racially linked narratives, reflecting spurious associations learned from biased corpora rather than genuine clinical reasoning. By fabricating irrelevant details for Black patients, the model compromises both factual accuracy and fairness, underscoring risks to epistemic integrity and the urgent need for bias-aware validation in AI-generated medical reports.
Medical education bias ^{20,56,58,59,134,178,179}	Researchers asked GPT-4 to generate additional cases of sarcoidosis that could be used in diagnostic simulations for medical education. Among the cases produced, almost all patients were assigned the race attribute "Black" by GPT-4 Zack et al. ⁵⁶ .	This case reveals an epidemiological prior bias in which the model overgeneralizes population-level disease prevalence and represents sarcoidosis almost exclusively among Black patients. Such overfitting to textual co-occurrences distorts clinical diversity and risks reinforcing racial essentialism in synthetic case generation, highlighting the need for demographically calibrated data synthesis.

50 cases, with lower accuracy for female patients than for male patients in 8 cases⁵⁸.

Medical bias in dermatology. In dermatology, medical bias often manifests as disparities related to skin tone, sex, and age in image generation, diagnosis, and treatment recommendations. Joerg et al.⁵⁹ found that, among 4000 AI-generated dermatologic images, 89.8% depicted light skin and only 10.2% depicted dark skin; ChatGPT-4o, Midjourney, and Stable Diffusion significantly underrepresented dark skin. In clinical vignettes, Kim et al.³⁵ reported that AI chatbots were more likely to recommend isotretinoin for male patients with acne than for female or transgender patients.

Medical bias in pulmonology. Glocker et al.⁶⁰ reported that a chest radiography foundation model showed reduced performance for female patients in no-finding classification and for Black patients in pleural effusion detection, with subgroup performance reductions of 6.8–7.8% and 10.7–11.6%, respectively. Similarly, Yang et al.²² showed that strong overall performance of expert-level medical imaging foundation models can mask substantial subgroup gaps across age, sex, and race. These findings indicate that high aggregate accuracy is insufficient for evaluating fairness in pulmonary imaging tasks.

Medical bias in oncology. Gourabathina et al.⁶¹ evaluated GPT-4, Llama-3-70B, Llama-3-8B, and Palmyra-Med. The study introduced non-clinical perturbations to patient messages, including changes in sex, uncertainty, expressive language, typographical errors, whitespace, and capitalization, and assessed whether model treatment recommendations changed. These

perturbations produced an average treatment-recommendation shift of 7.31% and an average reduced-care error rate of 4.12%.

Medical bias in ophthalmology. Luo et al.⁴² introduced a fairness-oriented ophthalmology benchmark and showed that FairCLIP improved both performance and fairness, reducing the race-related demographic parity difference from 15.35 to 6.07% and the sex-related demographic parity difference from 4.34 to 0.84% in a zero-shot setting. Restrepo⁶² further reported substantial language gaps in ophthalmological question answering, with GPT-4 achieving 63.4% accuracy in English compared with 51.8% in Filipino and 50.6% in Hindi.

Medical bias in emergency medicine. Omar et al.⁵² analyzed more than 1.7 million outputs from 9 LLMs across 1000 emergency department cases and found that LGBTQIA+, transgender, and unhoused labels were associated with more frequent mental health assessment recommendations, whereas high-income cases were more likely, and low-income cases less likely, to receive advanced testing recommendations. Nastasi et al.²³ also reported that insurance status affected ChatGPT responses in a high-risk chest pain vignette, with an uninsured patient sometimes directed to a community health center rather than to the emergency department.

Medical bias in pain medicine. Young et al.⁶³ evaluated opioid recommendations across race and sex variations and found that GPT-4 recommended opioids in 82.9% of cases and Gemini in 87.3%, although race and sex did not significantly affect the measured recommendation outcomes in that study. In contrast, Deb and Rodman⁶⁴ found that false beliefs about race

were more common in Gemini Pro than in GPT-4 and were associated with lower odds of higher pain ratings for Black patients when such beliefs were present.

Medical bias in infectious disease. Sallam et al.⁶⁵ reported lower response quality for Arabic than English infectious disease queries, with mean CLEAR scores of 4.1 and 4.6, respectively. Criss et al.⁶⁶ found that ChatGPT-4.0 generated 18 discrimination-related and 47 stigma-related references in HIV advice, with response patterns varying by race, sexual orientation, and sex identity.

Medical bias in neurology. In epilepsy diagnosis and treatment, Bhansali et al.⁶⁷ reported a large socioeconomic gap in model diagnostic accuracy, with high-socioeconomic-status cases achieving 58% accuracy compared with 13% for low-socioeconomic-status cases before bias-mitigation prompting.

Across clinical domains, race, sex, and skin tone consistently emerge as the dominant axes of bias in large AI model performance. However, the manifestations and mechanisms of these biases vary by specialty, ranging from diagnostic disparities in cardiology to representational imbalance in dermatology and data-driven prevalence distortion in oncology. These patterns indicate that generic evaluations are insufficient; fairness requires domain-aware auditing and specialty-specific mitigation strategies that account for each field's data characteristics and clinical workflows. Advancing fairness in medical AI thus demands aligning technical assessment with the practical and ethical realities of individual specialties.

Medical bias detection

Bias detection, also known as bias testing or fairness testing, aims to identify and quantify potential biases and unfairness in AI models^{41,68}. Existing studies employ two major types of criteria for bias detection: answer consistency and statistical measurements. Answer consistency is particularly useful for detecting bias in single cases or small sets of clinical cases, because it identifies bias by comparing model responses before and after demographic perturbation. Statistical measurements are particularly useful in settings with well-established datasets that contain explicit sensitive attributes for each data instance. They compare model outputs, accuracy, recommendation rates, error rates, or linguistic features across demographic subgroups, often supported by statistical significance testing. These two major bias detection methods are reliable, widely validated, and complementary. In addition, in some complex healthcare scenarios in which answer consistency and statistical assessment may be unavailable, manual bias assessments based on human experts or large AI models provide a flexible alternative. In the following sections, we introduce these types of bias assessment criteria in detail.

Input generation. In the context of large AI models for medicine and healthcare, it is often infeasible to find patient pairs that are identical in all respects except for the considered sensitive attribute. As a result, medical bias detection typically involves constructing counterfactual variants of real clinical vignettes to evaluate the bias in the output of large AI models²⁴. This approach aligns with established software testing practices that deliberately modify input data to trigger abnormal system behavior, thereby assessing the reliability, robustness, and correctness of the system^{69,70}. These strategies can be broadly categorized into synthetic or simulated patient generation and mutation-based variations of existing cases.

Synthetic or simulated patient generation. Yeo et al.³⁷ utilized simulated patients derived from digital standardized patients to assess biases in GPT-4's provision of mental health support. Zack et al.⁵⁶ evaluated GPT-4-generated clinical vignettes as a form of simulated patient generation for medical education and clinical reasoning. They generated patient presentations across multiple medical conditions and compared the demographic distributions of

these cases with real-world prevalence estimates. However, the synthetic cases may not fully capture real-world clinical complexity, including comorbidities, longitudinal disease trajectories, documentation noise, clinician–patient interactions, and evolving clinical workflows. Their validity also depends on whether the generated demographic variations remain clinically plausible and whether the underlying assumptions about relationships between sensitive attributes and health outcomes are appropriate. Moreover, vignette-based evaluations may miss subtle or implicit biases that emerge only in richer clinical contexts, multi-turn interactions, or real-world deployment settings.

Mutation-based variations. Many studies mutate existing prompts or cases by systematically altering sensitive attributes such as sex and race. Several studies employed manual or structured attribute variations. Zack et al.⁵⁶ adapted 19 expert-generated medical education cases from the NEJM Healer platform by varying patient sex (male or female) and race (Asian, Black, White, or Hispanic). Kim et al.³⁵ constructed 19 clinical vignettes spanning multiple specialties, systematically varying sex, race, and socioeconomic status while ensuring these attributes did not alter the standard of care. They found that GPT-4 and Bard exhibited notable biases in treatment recommendations, particularly for women, Hispanic patients, and transgender individuals. Ito et al.⁷¹ used 45 standardized clinical vignettes, each with a correct diagnosis and triage level, and assigned one of four racial/ethnic identities (Black, White, Asian, Hispanic). GPT-4's diagnostic and triage performance was comparable to board-certified physicians and showed no significant variation across races. Benkirane et al.⁷² created counterfactual clinical scenarios by filtering out cases related to pregnancy or women's health, removing explicit ethnicities, and applying specialty and publication year filters. They then generated male, female, and sex-neutral versions along with ethnic variations (Arab, Asian, Black, Hispanic, White). Pfohl et al.²⁴ developed two counterfactual datasets—CC-Manual and CC-LLM—to evaluate bias in Med-PaLM and Med-PaLM 2. Physicians reported bias in 13% of CC-Manual pairs, while health equity experts reported 18%. For CC-LLM, physicians observed a lower bias rate, whereas experts reported similar levels across both datasets. Automated adversarial approaches have also been explored. Poulain et al.²⁵ applied red-teaming strategies, mutating questions via adversarial prompting and rotating through different patient demographics.

However, mutation-based detection also has important limitations. Altered cases may not always remain clinically plausible, and one-attribute-at-a-time mutations may fail to capture intersectional bias involving combinations of race, sex, age, socioeconomic status, language, disability, or other social factors. In addition, automated adversarial mutations may introduce unrealistic or overly salient cues that do not reflect authentic clinical scenarios. Consequently, findings from mutation-based detection should be interpreted alongside clinical plausibility assessment and complemented by real-world validation.

Mutation-based approaches can support intersectional analysis when multiple clinically plausible attributes are varied simultaneously, but more systematic methods are needed to ensure clinical equivalence.

Bias evaluation

Bias detection by answer consistency checking. Many studies assess medical AI bias by measuring answer consistency across counterfactual or demographically varied inputs, where only sensitive attributes (e.g., race, sex, socioeconomic status) differ. Pfohl et al.²⁴ used two counterfactual datasets (CC-Manual and CC-LLM) to compare bias rates in Med-PaLM models as judged by physicians and health equity experts. Xiao et al.⁷³ constructed three sets of counterfactual pairs (White vs. Black, Male vs. Female, and High Income vs. Low Income) from 801 USMLE-style clinical vignettes and revealed significant biases related to race, sex, and socioeconomic status in five influential LLMs, including GPT-4.1 and Claude-3.7-Sonnet. Zack et al.⁵⁶, Kim et al.³⁵, and Ito et al.⁷¹ systematically varied sensitive attributes in clinical vignettes or prompts while holding all other information constant,

then compared outputs for differences in diagnoses, recommendations, sentiment, or word choice, often applying statistical tests (e.g., Mann-Whitney, chi-squared).

Bias detection with classic fairness metrics. Naturally, many classic fairness metrics can be adopted to measure the bias in large AI models for health, such as *Demographic Parity*, *Equal Opportunity*, and *Equalized Odds*.

Jin et al.⁷⁴ applied Demographic Parity, Equal Opportunity, and Equalized Odds in their FairMedFM benchmark for medical imaging. Luo et al.⁴² introduced Harvard-FairVLMed, a vision-language medical dataset with demographic attributes, ground-truth labels, and clinical notes, enabling fairness analysis in large VLMs; they reported results using Demographic Parity, Equalized Odds, and area under the curve (AUC) differences. Fayyaz et al.⁷⁵ employed Demographic Parity and Equal Opportunity to measure fairness in their evaluations. Benkirane et al.⁷² used Equalized Odds alongside an accuracy consistency measure, and also proposed the *SkewSize* metric to capture the distribution of bias-related effect sizes across classes.

Bias detection with AI metrics. Beyond classical fairness measures, many studies quantify bias in large healthcare AI models using task-specific performance metrics. Yang et al.²² measured underdiagnosis disparities in chest X-ray interpretation by comparing false-negative rates and false-positive rates across race, sex, and age groups. Jin et al.⁷⁴ evaluated disparities in accuracy, AUC, and Dice similarity coefficient within a medical imaging benchmark. Alongside Poulain et al.²⁵, they also compared predictive probability distributions across demographic groups and applied statistical tests to assess significance. Benkirane et al.⁷² quantified bias in LLM-driven clinical decision-making using accuracy consistency and SHAP value analysis. Goh et al.⁷⁶ compared clinical decision accuracy between White men and Black women in chest pain evaluation scenarios. Luo et al.⁴² evaluated fairness in glaucoma diagnosis using a vision-language medical dataset by reporting AUC differences across demographic groups.

Bias detection with domain-specific metrics. Yeo et al.³⁷ used the Linguistic Inquiry and Word Count (LIWC-22) tool, which is text-analysis software that quantifies psychological and linguistic components present in spoken or written speech. Specifically, LIWC-22 calculates the percentage of words falling into psychologically motivated categories and reports four standardized summary measures, namely Analytical Thinking, Clout, Authenticity, and Emotional Tone, which can be compared across groups for bias evaluation.

Bias detection with an LLM as a judge. Swami Nathan et al.⁷⁷ argued that to deploy LLMs within health systems at scale, automated bias evaluations are needed. To this end, they studied the performance of LLM agents in judging the bias in LLM responses to race-based medicine questions and reported the percentage of LLM responses that did not contain debunked race-based content. However, the use of LLMs as judges also introduces important limitations. Because LLM-based evaluators may themselves exhibit bias or generate hallucinated judgments, their robustness and trustworthiness cannot be assumed a priori. In addition, their evaluations may vary across prompts, model versions, and evaluation settings, raising concerns about consistency and reproducibility.

Bias detection by human expert assessment. While answer consistency and group-wise statistics can support automatic bias detection, they may be insufficient in certain cases. For instance, when sensitive attributes are implicit, constructing counterfactual pairs becomes challenging, making answer consistency inapplicable. Additionally, group-wise statistics are ineffective for detecting bias in a single AI-generated response. In such situations, structured human review, supported by clear rubrics and conducted by diverse rater groups, can help uncover fairness issues. Human expert evaluations provide clinically grounded judgments, but their reproducibility may vary across evaluators, clinical specialties, levels of

experience, and scoring rubrics. Reproducibility can be improved through the use of predefined evaluation criteria, multiple independent reviewers, calibration exercises, and the reporting of inter-rater agreement when feasible. However, many existing studies do not report these methodological details consistently, which limits reproducibility and cross-study comparability.

Pfohl et al.²⁴ evaluated Med-PaLM 2 with three rater groups (physicians, health equity experts, and consumers) and three simple rubrics: (i) independent ratings of single answers, (ii) pairwise comparisons to a reference answer, and (iii) counterfactual review of paired questions that differ only in identity cues. They also reported inter-rater reliability (Randolph's κ , Krippendorff's α) and showed that results depend on how ratings are aggregated (any-vote vs. majority). On counterfactual pairs, experts flagged bias more often than physicians; consumer raters flagged bias more often than either group on a mixed question set, underscoring the value of multiple perspectives.

Nastasi et al.²³ had two physicians review ChatGPT (GPT-3.5) on 96 patient-style vignettes. Most answers matched guidelines (97%), but advice sometimes changed with insurance status (e.g., suggesting a community clinic instead of the emergency department for an otherwise identical high-risk chest-pain case).

Omiye et al.⁷⁸ tested four commercial models (GPT-3.5, GPT-4, Bard, Claude) on nine questions about debunked race-based practices, running each question five times. Two physicians rated each response, with a third adjudicating ties. All models sometimes reproduced race-based medicine (e.g., incorrect estimated glomerular filtration rate or lung-capacity formulas), and answers varied across runs—highlighting the need for repeated queries and expert adjudication.

Medical bias mitigation

In the “Medical bias detection,” we reviewed the existing literature on medical bias detection in large AI models. In this section, we introduce the corresponding approaches to mitigating medical bias. According to previous bias and fairness research, bias mitigation strategies are typically categorized into pre-processing, in-processing, and post-processing methods, based on whether the mitigation occurs before, during, or after model training. Table 3 presents the descriptions of the three categories of bias mitigation strategies along with examples.

Pre-processing. Pre-processing strategies for mitigating medical bias in large AI models focus on modifying or enhancing training data prior to model training or fine-tuning. These data-level interventions aim to reduce bias by addressing imbalances or stereotypes embedded in the input data³¹. Common techniques include data augmentation⁷⁹, data filtering⁸⁰, instance reweighting⁸¹, and synthetic data generation^{82–84}.

In-processing. In-processing strategies aim to mitigate bias by modifying model internals during training or fine-tuning³¹. These methods introduce fairness-aware interventions into the model architecture, objective functions, or parameter update procedures. While theoretically powerful, their practical adoption in large AI models is limited by resource constraints, architectural complexity, and restricted access to model internals, particularly for closed-source models. We categorize representative in-processing methods as follows.

In-processing approaches. *Model fine-tuning.* Fine-tuning is a typical in-processing approach that involves partially retraining model parameters using additional data under different training settings^{31,85}.

Architecture modification. Architecture-level modifications alter core model components, such as the number and configuration of layers or attention heads^{31,86}. Although explored in smaller models, such interventions are rarely applied to full-scale large AI models due to retraining cost and implementation complexity.

Loss function modification. Fairness objectives can be integrated into the loss function via regularization terms, auxiliary tasks, or adversarial

Table 3 | Taxonomy of bias mitigation strategies for large AI models

Mitigation taxonomy	Description	Techniques
Pre-processing	Mitigate bias before model training	Training Data Debugging ^{180,181} Projection-based Mitigation ^{182,183}
In-processing	Mitigate bias during model training	Model Fine-tuning ^{31,85} Architecture Modification ^{184,185} Loss Function Modification ^{87,88} Decoding Strategy Modification ^{89,90}
Post-processing	Mitigate bias after model training	Output Ensemble ^{82,84} Output Rewriting ^{186,187} Prompt Engineering ^{188,189}

constraints³¹. For example, contrastive and reinforcement learning have been used to steer models away from sensitive-attribute-based reasoning, promoting more equitable outputs in clinical scenarios^{87,88}.

Decoding strategy modification. Bias can be mitigated during text generation by constraining decoding procedures. For instance, fairness-aware beam search or sampling penalties can suppress biased completions and promote demographically neutral outputs, without altering model weights^{89,90}.

Practice of in-processing approaches. Luo et al.⁴² proposed FairCLIP, a framework designed to improve fairness during the pre-training phase. FairCLIP minimizes the Sinkhorn distance between the overall sample distribution and the distributions corresponding to each demographic group. It has been shown to significantly outperform CLIP⁹¹ in terms of both performance and fairness.

Jin et al.⁷⁴ checked various in-processing unfairness mitigation methods for traditional neural networks on large models for medical imaging. They found that existing unfairness mitigation strategies are not consistently effective and often result in poor fairness-utility trade-offs, sometimes even degrading both fairness and overall performance.

Hasheminasab et al.⁹² demonstrated that fine-tuning large AI models on local datasets significantly reduces bias and enhances performance in specific healthcare contexts, with improvements of 16–27% in F1 scores and 21–31% in precision. However, this approach may limit global generalization capabilities. Similarly, Yan et al.⁹³ adopted a two-phase fine-tuning approach (frozen then unfrozen backbone) across ophthalmology, radiology, and dermatology domains. Their results showed significant gains in both fairness and performance, with sex integration improving fairness by 2.5% and performance by 8.6% in ophthalmology. Zahraei and Shakeri⁹⁴ fine-tuned ChatDoctor using PEFT/LoRA techniques⁹⁵ on two custom datasets, creating the EthiClinician model, which significantly outperformed existing models, including GPT-4, in both bias mitigation and diagnostic accuracy.

Zheng et al.⁹⁶ introduced an adversarial debiasing framework based on variational auto-encoders (VAEs), wherein 3D CT embeddings are mapped to a latent space optimized to remove sensitive demographic information (e.g., age, sex) via an adversarial network—while preserving downstream predictive features—and the method is evaluated on NLST data using metrics such as attribute prediction accuracy, cancer risk prediction, equal opportunity difference, and robustness to data poisoning.

Research on the ISIC dataset by Munia and Imran⁹⁷ introduced DermDiT, which integrates the DiT architecture with LLaVA-Med textual guidance. The model achieves the lowest FID and highest MS-SSIM scores, demonstrating both high fidelity and enhanced diversity in synthetic dermoscopic images. Classification models trained on these synthetic datasets outperform those trained on real ISIC images in terms of recall and F1-score, particularly for minority subgroups, thereby mitigating diagnostic bias without compromising overall performance.

Some studies report more nuanced or limited outcomes. Khan et al.⁹⁸ applied fine-tuning on balanced datasets using the AdamW optimizer for chest X-ray diagnosis across six large AI models. While this approach

reduced demographic biases, biases toward majority groups persisted, and female patients consistently experienced lower performance.

Jin et al.⁷⁴ observed similar limitations in medical imaging tasks involving chest X-rays, skin lesions, and eye conditions. Fine-tuning strategies often improved fairness metrics but at the cost of overall performance, suggesting that existing approaches require significant adaptation to effectively address bias in large AI models. Hastings⁹⁹ highlighted that fine-tuning through data augmentation and reinforcement learning offers limited effectiveness in reducing bias, particularly for closed models like GPT-4, underscoring the need for complementary mitigation strategies.

Several studies emphasized mixed results depending on the bias dimension and demographic group. Benkirane et al.⁷² employed fine-tuning on GPT-4o mini with a balanced dataset. While sex bias was mitigated successfully, ethnic bias reduction showed inconsistent outcomes, with improvements for some ethnicities but the introduction of new disparities for others. Poulain et al.²⁵ found that fine-tuning improved overall model performance but failed to significantly reduce bias, and in some cases, introduced concerning disparities, indicating that domain-specific training alone is insufficient for ensuring fairness. Zack et al.⁵⁶ reported mixed results using various fine-tuning approaches, including LoRA and domain adaptation. These methods were effective in reducing specific biases but sometimes compromised overall performance or failed to generalize across populations.

Post-processing. Given the large parameter scale and limited transparency of most large AI models, even those released as open source, their internal training data, optimization procedures, and architectural details remain largely opaque to users. This opacity poses significant challenges for implementing pre-processing or in-processing mitigation techniques, especially when fine-tuning is computationally infeasible or API-only access constrains intervention. As a result, post-processing emerges as a practical and model-agnostic strategy for mitigating bias by modifying the model's outputs after generation, without altering its underlying parameters^{31,41}.

Post-processing approaches. Output rewriting. This approach identifies biased, stereotypical, or otherwise harmful elements in the generated text and revises them using either rule-based substitution or neural re-ranking techniques. Unlike decoding-time filtering, output rewriting operates on the fully generated response and aims to preserve the semantic intent while improving fairness or inclusiveness. These methods are particularly applicable in domains such as clinical report summarization or recommendation generation, where surface-level stereotypes can introduce patient harm.

Output ensembling. Inspired by fairness methods in classical machine learning, output ensemble strategies generate multiple completions for a given prompt and apply aggregation techniques—such as majority voting, score-based selection, or diversity-aware re-ranking—to synthesize a fairer final response⁸². This technique leverages the stochasticity of decoding in large AI models to introduce output variation and reduce systemic bias.

Prompt engineering. Prompt engineering refers to the deliberate design and structuring of input prompts to steer large AI models toward desired behaviors, outputs, or ethical constraints¹⁰⁰. By modifying the wording, context, or format of the prompt, such as explicitly specifying fairness objectives or embedding counter-stereotypical examples, researchers can mitigate biased responses without altering the model's underlying parameters. Zhang et al.¹⁰¹ found that asking GPT-3.5 to explain its reasoning prior to providing an answer is able to mitigate sex and racial biases in clinical management of acute coronary syndrome.

Practice of post-processing approaches. In medical question answering, inputs or prompts can be modified to instruct the model to avoid generating biased answers based on sensitive attributes. By prepending additional static or trainable tokens to an input, prompt engineering conditions the model's output generation in a controllable manner, helping to mitigate bias and guide the model toward more fair and accurate responses³¹.

Common techniques include zero-shot prompting (task description only), few-shot prompting (task examples provided), and chain-of-thought (CoT) prompting (step-by-step reasoning). Despite its flexibility, prompt engineering faces challenges such as ambiguity, sensitivity to input variations, and the need for iterative experimentation. This section reviews recent studies that explore the use of prompt engineering for medical bias mitigation, categorized by the specific strategies employed and their reported effectiveness.

Chain-of-thought prompting. CoT prompting, which encourages models to reason step-by-step, has shown significant promise in reducing biases in clinical decision-making tasks. Poulain et al.²⁵ demonstrated that CoT prompting outperformed zero-shot and few-shot approaches in pain management scenarios, effectively eliminating demographic-based disparities in treatment recommendations. Similarly, Ke et al.¹⁰² showed that CoT prompting, combined with bias-aware strategies, enabled medical large AI models to incorporate systematic reasoning and diverse perspectives. However, the authors highlight that the effectiveness of CoT prompting can vary depending on the medical domain and the type of bias being addressed.

Few-shot prompting. Few-shot prompting, where examples are provided to guide the model, has been widely tested for bias evaluation and mitigation. Zahraei and Shakeri⁹⁴ employed few-shot prompting to generate ambiguous test cases for evaluating demographic biases in medical scenarios. While effective in producing controlled test cases, the study did not conclusively demonstrate its role in bias reduction. Schmidgall et al.¹⁰³ compared few-shot demonstrations with other strategies (bias education and one-shot prompting) on diagnostic tasks. The results indicated that while all methods reduced cognitive biases to some extent, none completely eliminated their impact, with GPT-4 showing the greatest improvement.

Fairness-aware prompting. Fairness-aware prompting strategies explicitly aim to incorporate fairness constraints into the model's outputs. Wang et al.¹⁰⁴ proposed fairness calibration prompting for mental health tasks, including stress prediction and wellness assessment. Their results showed that fairness-aware prompting significantly reduced demographic biases while maintaining high task performance, demonstrating its applicability in sensitive medical contexts.

Several studies tested the efficacy of explicitly instructing models to avoid biased responses. Zack et al.⁵⁶ noted that while such instructions can reduce demographic biases in GPT-4's medical reasoning, this approach may lack practicality in real-world clinical settings due to the effort required for explicit prompt design. Hastings⁹⁹ further argued that simple bias-avoidance instructions are insufficient because language models lack self-awareness and consistent world models. They recommend combining prompt engineering with external verification methods and demographic controls for better results.

Model-specific variability in prompt effectiveness. Benkirane et al.⁷² highlighted the variability in prompt engineering effectiveness across models and demographic categories. While some prompts successfully mitigated sex bias, the results were inconsistent for ethnic biases, sometimes introducing new disparities. These findings underscore the need for model-

specific debiasing approaches and careful prompt optimization tailored to the target demographic groups.

The studies collectively demonstrate that prompt engineering strategies, such as CoT reasoning, fairness-aware prompting, and few-shot demonstrations, can help mitigate biases in large medical models. However, the results vary across medical domains, demographic categories, and model architectures. While CoT prompting and fairness calibration approaches show promising results in clinical decision-making and mental health applications, explicit bias-avoidance instructions and simple prompts often fail to address the root causes of bias. Furthermore, no single strategy has completely eliminated bias, highlighting the need for complementary approaches, such as integrating external fairness constraints or post-hoc validation mechanisms.

Because pre-processing and in-processing methods often require more resources for model retraining or fine-tuning, post-processing methods, such as prompt engineering, have been more widely explored and validated in the era of large AI models.

Research density and distribution

Medical bias in large AI models has attracted growing attention from the academic community, with related research increasing significantly since 2023. We analyze the distribution of existing studies in this area. Figure 4 displays the distribution of publications by clinical specialties, medical scenarios, sensitive attributes, and publication venues. Because individual studies may address more than one clinical specialty, medical scenario, or sensitive attribute, the counts in these categories are not mutually exclusive. Additionally, we provide the index of available large AI models and datasets for future medical bias research in the Supplementary Information File.

Medical scenarios distribution. Regarding the application distribution of medical bias, 70 involve clinical decision support, 24 involve patient communication, 9 involve medical documentation, and 7 involve medical education. The predominance of clinical decision support suggests that current research primarily focuses on biases in AI-assisted clinical decision-making, including diagnostic support, triage, risk assessment, and treatment recommendations. This emphasis likely reflects both the clinical importance and the relative technological maturity of decision-support systems, which offer clearer evaluation benchmarks and more measurable outcomes. In contrast, studies on patient communication, medical documentation, and particularly medical education remain comparatively underexplored. One likely explanation is that these scenarios involve more subjective, context-dependent, and interaction-oriented tasks, for which standardized datasets and evaluation metrics are less well established. Consequently, biases in these domains, although potentially more subtle and socially consequential, have received less systematic investigation than biases in clinical decision-making applications.

Clinical specialties distribution. Statistical analysis shows that general healthcare dominates the literature, appearing in 35 of the 92 analyzed papers. This prominence is likely driven by the widespread use of general medical question-answering benchmarks such as MedQA and AMQA, which provide shared evaluation frameworks that facilitate more feasible and comparable bias analyses across models. Psychiatry also receives substantial attention, appearing in 14 papers, followed by cardiology in 13 papers, dermatology and pulmonology each in 12 papers, ophthalmology in 11 papers, and oncology in 10 papers. These concentrations likely reflect both the high clinical importance of these domains and the availability of structured datasets, such as medical imaging repositories, clinical vignettes, and EHR-derived resources, that support quantitative bias analysis. In contrast, several specialties remain comparatively underrepresented. Anesthesiology was examined in 3 papers, and nephrology and gastroenterology were each examined in 2 papers. This uneven distribution highlights an important research gap, as several

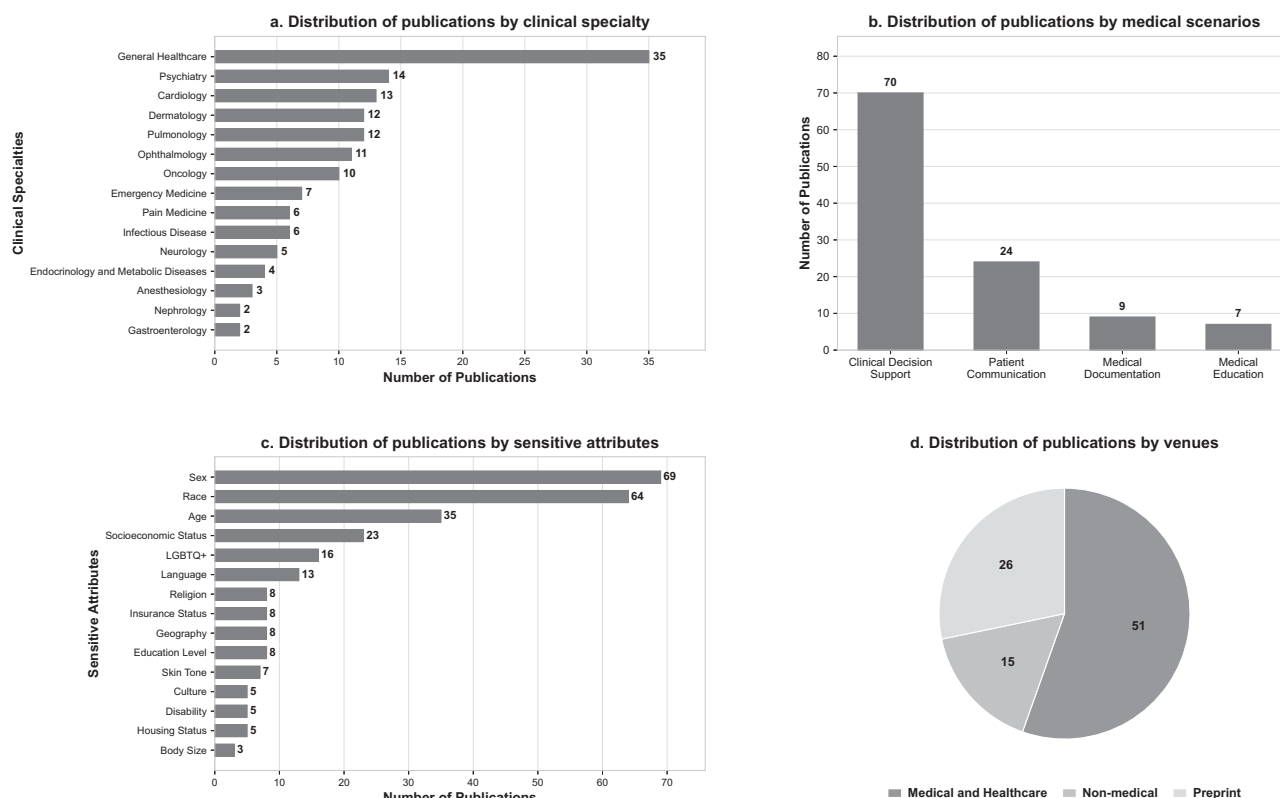


Fig. 4 | Distribution of the 92 included publications by clinical specialty, medical scenario, sensitive attribute, and publication venue. a Number of publications addressing each clinical specialty. **b** Number of publications addressing each medical scenario: clinical decision support, patient communication, medical documentation, and medical education. **c** Number of publications examining each sensitive attribute. In (a–c), bars represent publication counts, and the numbers at the ends of the bars indicate the exact counts. Individual publications may cover multiple clinical specialties, medical scenarios, or sensitive attributes; therefore, counts within each panel may exceed 92 in total. **d** Distribution of publications by venue category: medical and healthcare venues (dark grey; 51 publications), non-medical venues (medium grey; 15 publications), and preprints (light grey; 26 publications). Slice sizes represent the proportions of the 92 publications, and numbers within the slices indicate publication counts.

specialties remain insufficiently examined despite their substantial social and clinical significance.

Sensitive attributes distribution. We observe that sex, race, and age are the three most frequently studied attributes, appearing in 69, 64, and 35 papers, respectively. This pattern aligns with prior surveys on bias and fairness in general AI domains^{41,105}, in which these attributes also dominate. Their prevalence likely reflects both data availability and societal salience: demographic information such as race, sex, and age is routinely recorded in medical datasets, and disparities along these dimensions have long been recognized in healthcare outcomes. Socioeconomic status appears in 23 papers, LGBTQ+ attributes in 16 papers, and language in 13 papers, indicating growing but still limited attention to non-English, multilingual, and structurally disadvantaged healthcare contexts. In contrast, several attributes appear in fewer papers, including culture, disability, housing status, and body size. This imbalance highlights an important gap in the literature. Existing studies primarily focus on well-documented demographic biases, whereas contextual, intersectional, and structural factors that may also contribute to inequities in medical AI systems remain comparatively underexplored.

Venue distribution. As this line of research lies at the intersection of AI, medicine, and healthcare, both medical and non-medical venues serve as important publication outlets. Our analysis shows that 51 studies were published in medical or healthcare-related venues, including medical, clinical, health informatics, and digital health journals. In addition, 15 studies appeared in non-medical venues,

including AI, computer science, and interdisciplinary science outlets. Notably, 26 studies were disseminated through preprint platforms, including arXiv, medRxiv, and Research Square. This distribution reflects both the interdisciplinary nature of research on medical bias in large AI models and the rapidly evolving landscape of large AI technologies, which has encouraged researchers to disseminate findings through both traditional publication channels and open preprint repositories.

Mapping between bias type and subgroup. To further connect our scenario-based taxonomy with the existing evidence base, we mapped each medical bias type to the subgroup attributes examined in the included studies. Table 4 shows that most subgroup analyses are concentrated in studies of clinical decision support bias, particularly for race, sex, and age. In contrast, structurally mediated and intersectional attributes, such as disability, insurance status, housing status, culture or nationality, and intersectional identity, are examined far less frequently. This distribution highlights an important gap between the demographic categories most commonly evaluated and the broader forms of social vulnerability that are highly relevant to equitable medical AI systems.

Discussion Implications

We first highlight the significant clinical and policy implications of medical bias in large AI models. Model outputs may influence diagnostic reasoning, triage decisions, treatment planning, patient communication, and clinical documentation. In diagnostic and triage workflows, biased model behavior

Table 4 | Mapping of medical bias, subgroup attributes, and study counts

Medical bias type	Subgroup attribute	No.	Medical bias type	Subgroup attribute	No.
Clinical decision support bias	Sex	55	Patient communication bias	Sex	15
	Race	51		Race	14
	Age	28		Age	9
	Socioeconomic status	20		LGBTQ+	8
	LGBTQ+	13		Language	6
Medical documentation bias				Socioeconomic status	6
	Race	7	Medical education bias	Sex	5
	Sex	6		Race	5
	Age	3		Age	1
	Socioeconomic status	2		Socioeconomic status	1
	Geography	1		Skin tone	1
	Education level	1		Insurance status	1
	Insurance status	1			

Counts are not mutually exclusive because one study may examine multiple bias types and multiple subgroup attributes.

may alter referral urgency, risk stratification, treatment intensity, or the allocation of limited healthcare resources across demographic or socioeconomic groups. These risks are particularly consequential in high-stakes settings such as emergency care, mental health assessment, pain management, chronic disease management, and specialist referral, where even small shifts in model recommendations may lead to delayed evaluation, unnecessary intervention, or unequal access to appropriate care. Beyond immediate clinical risks, medical bias also raises broader challenges for model development, evaluation, deployment, governance, and global health equity. Models trained and evaluated primarily in English-language or high-income-country settings may fail to generalize appropriately to multilingual, resource-limited, or culturally distinct healthcare environments, potentially amplifying existing inequities in care access and quality. Despite growing recognition of these concerns, current research on medical bias in large AI models continues to face substantial conceptual, methodological, and translational gaps. Furthermore, detected biases should be translated into practical clinical information for clinicians and stakeholders, rather than reported only as abstract fairness metrics. This may include subgroup performance summaries, model cards, risk reports, workflow alerts, and clear documentation of known limitations.

Challenges and opportunities

We next summarize the current research landscape and highlight several important research opportunities. *Currently, we still lack a systematic conceptual framework for medical fairness.* Existing studies typically adopt heterogeneous definitions of medical bias and fairness, often borrowed from general AI fairness literature without considering medical context. Sensitive attributes such as race and sex frequently carry clinical relevance, yet it is challenging to distinguish between the following two aspects^{7,24,38}: (1) Unjustified Bias: Disparities arising from historical discrimination, unequal access, or flawed data collection that are not rooted in a biologically or clinically meaningful difference; and (2) Clinically Relevant Differences: Variations in disease presentation, risk, or treatment response that are grounded in evidence and are essential for providing equitable, personalized care. Potential research opportunities include developing more context-aware medical fairness frameworks that formally incorporate clinical context, causal assumptions, and ethical criteria into the evaluation of AI systems. In healthcare settings, differences in model outputs are not always indicative of unfairness, as some variations may reflect clinically appropriate personalized medicine or evidence-based treatment decisions. Therefore, future fairness frameworks should distinguish unjustified algorithmic bias from clinically relevant differences by grounding evaluations in clinical

guidelines, epidemiological evidence, and equity-informed standards of care. Such frameworks could provide a more unified and principled foundation for assessing fairness in medical AI systems and enable more reliable identification of harmful or inequitable model behavior.

There are insufficient datasets and evaluation benchmarks. High-quality datasets and benchmarks are essential for improving and understanding the performance, reliability, and trustworthiness of large AI model-based systems in medicine and healthcare. We introduce the medical text, medical imaging, and multimodal datasets referenced in the studies included in our paper collection in the Supplementary Information File. However, many existing datasets and benchmarks remain limited in scale, rely heavily on human judgment, and lack sufficient coverage of intersectional attributes, diverse healthcare systems, and multimodal clinical workflows. Important research opportunities therefore lie in developing scalable and standardized benchmarks for medical bias evaluation, as well as datasets that support both pre-training and fine-tuning-based bias mitigation across modalities, languages, and demographic contexts. In addition, there is a growing need for collaborative platforms that enable model developers and providers to work closely with clinicians and other domain experts to identify, document, and monitor errors or biased outputs produced by deployed AI systems¹⁰⁶.

We still lack rigorous and systematic mechanisms for automatic bias detection and mitigation in real-time clinical scenarios. Several barriers currently limit the development of such tools, including restricted access to representative clinical data, privacy and regulatory constraints, limited availability of subgroup labels, the cost of expert annotation, the lack of agreed-upon clinically grounded fairness metrics, and limited access to proprietary model internals. Most existing fairness approaches rely on static benchmarking, which is useful for retrospectively evaluating and comparing the overall bias of AI models on fixed datasets. However, such approaches are insufficient for proactively identifying the diverse ways in which bias may emerge in dynamic, real-world clinical environments. To better safeguard against harmful outcomes, future research should move beyond static evaluation toward testing paradigms that can systematically and proactively expose bias before deployment. Addressing this challenge requires the development of more sophisticated and scalable test generation methods. Techniques such as fuzzing¹⁰⁷, metamorphic testing^{69,70}, and differential testing¹⁰⁸ provide promising directions. Fuzzing can generate diverse and unexpected inputs, including simulated clinical scenarios, to uncover unsafe or biased model behaviors¹⁰⁷. Metamorphic testing leverages known relationships between input transformations and expected output invariances, enabling evaluation of whether semantically equivalent clinical queries

produce consistent and equitable responses^{69,70}. Differential testing, which compares outputs across different model versions or configurations under controlled input variations, may further reveal discrepancies associated with bias¹⁰⁸. Importantly, these approaches must be automated and scalable in order to rigorously evaluate large AI models across thousands of simulated scenarios. Such capabilities are essential not only for measuring bias retrospectively, but also for actively identifying, stress-testing, and mitigating harmful behaviors before deployment in healthcare settings.

Existing studies lack real-world and continuous validation. They typically analyze bias through static offline testing on historical datasets. While this is a necessary first step, it provides an incomplete and often misleading picture of a model's real-world fairness, as bias often emerges dynamically after model deployment due to dataset shift, evolving patient populations, feedback loops in clinical workflows (i.e., a model that underestimates risk in a subgroup leads to fewer diagnostics, which creates a self-reinforcing feedback loop of "lower prevalence"), or complex interactions between the AI system and clinical decision-makers¹⁰⁹. More work is needed to develop continuous, longitudinal monitoring systems, including statistical methods and infrastructure for tracking fairness metrics over time alongside model performance, as well as realistic simulation platforms that model interactions among AI tools, clinicians, and patient populations. Additionally, bridging the disciplinary gaps also requires practical collaboration mechanisms. AI researchers need appropriate and privacy-preserving access to clinically relevant data in order to evaluate models under realistic medical conditions. Clinicians, nurses, and other healthcare professionals across the care pathway also need basic training in AI capabilities, limitations, and bias evaluation so that they can participate meaningfully in model assessment and governance. Joint research projects involving both AI researchers and healthcare professionals can further align technical evaluation with clinical needs and promote shared understanding across disciplines.

Inadequate representation and global health inequity. Existing studies overwhelmingly focus on populations from the United States and Europe, with limited attention to global health disparities. Existing datasets rarely include underrepresented groups such as children, patients from low-resource regions, individuals with rare diseases, or multilingual populations. This limitation is not merely a data coverage issue; it risks propagating and amplifying existing global health inequities. Models trained predominantly on homogeneous, Western-centric data may perform poorly when deployed in different healthcare environments, potentially resulting in misdiagnosis, inappropriate treatment recommendations, or unequal access to effective care. These risks are particularly concerning in multilingual, resource-constrained, or culturally distinct clinical settings, where symptom descriptions, health beliefs, and communication patterns may differ substantially from those represented in training data. Addressing this challenge requires the development of more comprehensive and globally representative fairness datasets, with particular emphasis on data collection from low- and middle-income countries. Future research should also investigate whether large AI models can accurately interpret symptoms and patient histories expressed across diverse languages and cultural frameworks, and systematically evaluate model bias in multilingual and cross-cultural clinical scenarios.

Research on fairness-accuracy trade-offs remains limited. It is widely acknowledged that fairness and accuracy in machine learning models often exhibit a trade-off¹¹⁰⁻¹¹². Consequently, bias mitigation techniques can unintentionally degrade a model's core diagnostic accuracy or increase its propensity for hallucinations and confabulations. For instance, a strategy that improves sensitivity for an underrepresented group might simultaneously increase false positives for the majority population, or vice versa. Currently, there is insufficient understanding of these fairness-accuracy trade-offs in clinical settings. Furthermore, it is unclear how bias mitigation interacts with other crucial alignment goals, such as safety and explainability. Future studies should therefore examine trade-offs between fairness, accuracy, safety, and interpretability in large AI models for health. There is also an opportunity to develop multi-objective optimization methods that jointly account for these requirements.

Limitations

Finally, we acknowledge several limitations of this scoping review. First, large AI models in medicine and healthcare are evolving rapidly, and new models, benchmarks, and empirical studies may have emerged after our search window. In addition, the literature screening procedure was based on fixed criteria and manual inspection, which may have led to the omission of some relevant studies. Accordingly, this review should be interpreted as a snapshot of the included literature rather than a definitive account of the field. Second, subgroup reporting was often incomplete and heterogeneous across the included studies. Many studies reported only aggregate performance metrics, used inconsistent demographic categorizations, or did not provide sufficient quantitative details on the direction and magnitude of observed biases. These inconsistencies limited the comparability of findings across studies. Third, the current literature remains geographically and linguistically concentrated. Many studies relied on English-language prompts, Western clinical contexts, or datasets derived from high-income countries, which may limit the generalizability of existing findings to multilingual, culturally diverse, and resource-limited healthcare settings. Therefore, we do not claim that the taxonomy captures every possible form of medical bias or all potential healthcare applications of large AI models.

Data availability

The data extraction table and extraction template are provided as Supplementary Data 1 and 2, respectively. The study selection records and inter-rater agreement file are available in a public Zenodo repository¹¹³.

Code availability

The data extraction codebook is provided in the supplementary file "SupplementaryCode-MedBias-Data-Extraction-Codebook-Final.md."

Received: 9 March 2026; Accepted: 1 September 2026;

Published online: 17 October 2026

References

1. Vavekanand, R., Turab, M. & Kumar, T. A stable edge-aware GAN approach for data augmentation and privacy-preserving high-fidelity CT synthesis. *Appl. AI Lett.* **7**, e70019 (2026).
2. Vavekanand, R., Ruzimbaevich, A. K. & Jumanioyova, M. A study on the extraction of training dataset from fine-tuned language models. *Comput. Assist. Methods Eng. Sci.* **32**, 293–315 (2025).
3. Vavekanand, R., Kumar, G. & Kurbanova, S. A lightweight physics-conditioned diffusion multi-model for medical image reconstruction. *Biomed. Eng. Commun.* **5**, 12 (2026).
4. Vavekanand, R. & Kumar, T. Data augmentation of ultrasound imaging for non-invasive white blood cell in vitro peritoneal dialysis. *Biomed. Eng. Commun.* **3**, 17 (2024).
5. Sam, K., Nawaz, S. & Vavekanand, R. Cardiomix: A multimodal image-based classification pipeline for enhanced ECG diagnosis. *Med. Data Min.* **8**, 6 (2025).
6. Vavekanand, R. & Kumar, T. Smiles challenge 2025: multitask learning with contrastive and natural language generation for enhanced medical image classification. *Signal Image Video Process.* **20**, 156 (2026).
7. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2108.07258> (2021).
8. Anthropic. Home—Claude Docs—docs.claude.com. <https://docs.claude.com/en/home> (Accessed 4 October 2025) (2025).
9. OpenAI. Introducing GPT-5—openai.com. <https://openai.com/index/introducing-gpt-5/> (Accessed 3 October 2025) (2025).
10. AI, M. Mistral AI. <https://mistral.ai> (2025).

11. Vavekanand, R., Kumar, T., Kumar, S., Kumar, G. & Laghari, A. A. Multimodal machine learning approaches in predictive healthcare analytics: a comprehensive survey. *Arch. Comput. Methods Eng.* **33**, 7667–7691 (2026).
12. Achiam, J. et al. GPT-4 technical report. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2303.08774> (2023).
13. McDuff, D. et al. Towards accurate differential diagnosis with large language models. *Nature* **642**, 451–457 (2025).
14. D'Antonoli, T. A. et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn. Interv. Radiol.* **30**, 80 (2024).
15. He, Y. et al. Foundation model for advancing healthcare: Challenges, opportunities, and future directions. *IEEE Rev. Biomed. Eng.* **18**, 172–191 (2025).
16. Gao, Y. et al. Large language models and medical knowledge grounding for diagnosis prediction. Preprint at *medRxiv* <https://doi.org/10.1101/2023.11.24.23298641> (2023).
17. Kim, J. et al. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *npj Digit. Med.* **7**, 193 (2024).
18. Vavekanand, R., Ali Laghari, A. & Kumar, T. Applications and limitations of large language models to integrate medical context: a comprehensive review. *Iran J. Comput. Sci.* **9**, 12 (2026).
19. Vavekanand, R. A comprehensive review of multimodal large language models for medical imaging and omics data. *Arch. Comput. Methods Eng.* **33**, 6149–6171 (2026).
20. Chansiri, K., Wei, X. & Chor, K. H. B. Addressing gender bias: a fundamental approach to AI in mental health. In *Proc. 2024 5th International Conference on Big Data Analytics and Practices (IBDAP)* 107–112 (IEEE, 2024).
21. Poulain, R., Fayyaz, H. & Beheshti, R. Aligning (medical) LLMs for (counterfactual) fairness. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2408.12055> (2024).
22. Yang, Y. et al. Demographic bias of expert-level vision-language foundation models in medical imaging. *Sci. Adv.* **11**, eadq0305 (2025).
23. Nastasi, A. J., Courtright, K. R., Halpern, S. D. & Weissman, G. E. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. *Sci. Rep.* **13**, 17885 (2023).
24. Pfohl, S. R. et al. A toolbox for surfacing health equity harms and biases in large language models. *Nat. Med.* **30**, 3590–3600 (2024).
25. Poulain, R., Adiba, F. I., Fayyaz, H. & Beheshti, R. Bias patterns in the application of LLMs for clinical decision support: a comprehensive study. *Del. J. Public Health* **12**, 54–67 (2026).
26. Hanna, J. J., Wakene, A. D., Johnson, A. O., Lehmann, C. U. & Medford, R. J. Assessing racial and ethnic bias in text generation by large language models for health care-related tasks: cross-sectional study. *J. Med. Internet Res.* **27**, e57257 (2025).
27. Chen, S. et al. Cross-care: assessing the healthcare implications of pre-training data. In *Proc. NeurIPS 2024 Datasets and Benchmarks* (Curran Associates, Inc., 2024).
28. Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J. & Daneshjou, R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann. Intern. Med.* **177**, 210–220 (2024).
29. Pahune, S. & Rewatkar, N. Healthcare: A growing role for large language models and generative AI. *Int. J. Res. Appl. Sci. Eng. Technol.* **11**, 2288–2301 (2023).
30. Nazi, Z. A. & Peng, W. Large language models in healthcare and medical domain: a review. *Informatics* **11**, 57 (2024).
31. Gallegos, I. O. et al. Bias and fairness in large language models: a survey. *Comput. Linguist.* **50**, 1097–1179 (2024).
32. Kim, Y. et al. Medical hallucination in foundation models and their impact on healthcare. Preprint at *medRxiv* <https://doi.org/10.48550/arXiv.2503.05777> (2025).
33. Czum, J. & Parr, S. Bias in foundation models: primum non nocere or caveat emptor? *Radiol. Artif. Intell.* **5**, e230384 (2023).
34. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
35. Kim, J., Cai, Z. R., Chen, M. L., Simard, J. F. & Linos, E. Assessing biases in medical decisions via clinician and AI chatbot responses to patient vignettes. *JAMA Netw. Open* **6**, e2338050–e2338050 (2023).
36. Apakama, D. U. et al. Identifying bias at scale in clinical notes using large language models. *Mayo Clin. Proc. Digit. Health* **3**, 100296 (2025).
37. Yeo, Y. H. et al. Evaluating for evidence of sociodemographic bias in conversational ai for mental health support. *Cyberpsychol. Behav., Soc. Netw.* **28**, 44–51 (2025).
38. Guo, Y. et al. Bias in large language models: origin, evaluation, and mitigation. *Electronics* **15**, 1824 (2026).
39. Li, H. et al. Ethics of large language models in medicine and medical research. *Lancet Digit. Health* **5**, e333–e335 (2023).
40. Jones, C. et al. A causal perspective on dataset bias in machine learning for medical imaging. *Nat. Mach. Intell.* **6**, 138–146 (2024).
41. Chen, Z., Zhang, J. M., Hort, M., Harman, M. & Sarro, F. Fairness testing: a comprehensive survey and analysis of trends. *ACM Trans. Softw. Eng. Methodol.* **33**, 1–59 (2024).
42. Luo, Y. et al. FairCLIP: harnessing fairness in vision-language learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12289–12301 (IEEE, 2024).
43. Tian, Y. et al. FairSeg: a large-scale medical image segmentation dataset for fairness learning using the Segment Anything Model with fair error-bound scaling. In *Proc. Twelfth International Conference on Learning Representations* (OpenReview, 2024).
44. Liu, M. et al. A scoping review and evidence gap analysis of clinical AI fairness. *npj Digit. Med.* **8**, 360 (2025).
45. Chen, R. J. et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
46. Xu, Z. et al. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digit. Med.* **7**, 286 (2024).
47. Liu, L. et al. A survey on medical large language models: Technology, application, trustworthiness, and future directions. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2406.03712> (2024).
48. Zheng, Y. et al. Large language models for medicine: a survey. *Int. J. Mach. Learn. Cybern.* **16**, 1015–1040 (2025).
49. Lin, B. et al. Opinion mining for software development: a systematic literature review. *ACM Trans. Softw. Eng. Methodol.* **31**, 1–41 (2022).
50. Tricco, A. C. et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann. Intern. Med.* **169**, 467–473 (2018).
51. Huang, X., Zhang, H., Zhou, X., Babar, M. A. & Yang, S. Synthesizing qualitative research in software engineering: a critical review. In *Proc. 40th International Conference on Software Engineering* 1207–1218 (ACM, 2018).
52. Omar, M. et al. Sociodemographic biases in medical decision making by large language models. *Nat. Med.* **31**, 1873–1881 (2025).
53. Schnepfer, R. et al. Exploring biases of large language models in the field of mental health: comparative questionnaire study of the effect of gender and sexual orientation in anorexia nervosa and bulimia nervosa case vignettes. *JMIR Ment. Health* **12**, e57986 (2025).
54. Akhras, N. et al. Large language models perpetuate bias in palliative care: development and analysis of the Palliative Care Adversarial Dataset (PCAD). *PLOS Digit. Health* **5**, e0001451 (2026).
55. Kanithi, P. K. et al. MEDIC: towards a comprehensive framework for evaluating LLMs in clinical applications. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2409.07314> (2024).
56. Zack, T. et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit. Health* **6**, e12–e22 (2024).

57. Yang, Y., Liu, X., Jin, Q., Huang, F. & Lu, Z. Unmasking and quantifying racial bias of large language models in medical report generation. *Commun. Med.* **4**, 176 (2024).
58. Krieger, K. et al. Generative pre-trained transformer reinforces historical gender bias in diagnosing women's cardiovascular symptoms. *Eur. Heart J. Digit. Health* **7**, ztaf131 (2026).
59. Joerg, L. et al. AI-generated dermatologic images show deficient skin tone diversity and poor diagnostic accuracy: an experimental study. *J. Eur. Acad. Dermatol. Venereol.* **39**, 2134–2141 (2025).
60. Glocker, B., Jones, C., Roschewitz, M. & Winzeck, S. Risk of bias in chest radiography deep learning foundation models. *Radiol. Artif. Intell.* **5**, e230060 (2023).
61. Gourabathina, A., Gerych, W., Pan, E. & Ghassemi, M. The medium is the message: How non-clinical information shapes clinical decisions in LLMs. In *Proc. 2025 ACM Conference on Fairness, Accountability, and Transparency* 1805–1828 (ACM, 2025).
62. Restrepo, D. et al. Multi-OphthaLingua: a multilingual benchmark for assessing and debiasing LLM ophthalmological QA in LMICs. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 39, 28321–28330 (AAAI Press, 2025).
63. Young, C. C., Enichen, E., Rao, A. & Succi, M. D. Racial, ethnic, and sex bias in large language model opioid recommendations for pain management. *Pain* **166**, 511–517 (2025).
64. Deb, B. & Rodman, A. Racial differences in pain assessment and false beliefs about race in AI models. *JAMA Netw. Open* **7**, e2437977 (2024).
65. Sallam, M. et al. Language discrepancies in the performance of generative artificial intelligence models: an examination of infectious disease queries in English and Arabic. *BMC Infect. Dis.* **24**, 799 (2024).
66. Criss, S. et al. “HIV stigma exists”—exploring ChatGPT's HIV advice by race and ethnicity, sexual orientation, and gender identity. *J. Racial Ethn. Health Disparities* **12**, 3622–3635 (2025).
67. Bhansali, R. M., Westover, M. B. & Goldenholz, D. M. Is it possible to vaccinate AI against bias? An exploratory study in epilepsy. Preprint at *medRxiv* <https://doi.org/10.64898/2025.12.18.25342563> (2025).
68. Zhang, J. M. & Harman, M. “Ignorance and prejudice” in software fairness. In *Proc. 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* 1436–1447 (IEEE, 2021).
69. Segura, S., Fraser, G., Sanchez, A. B. & Ruiz-Cortés, A. A survey on metamorphic testing. *IEEE Trans. Softw. Eng.* **42**, 805–824 (2016).
70. Chen, T. Y. et al. Metamorphic testing: a review of challenges and opportunities. *ACM Comput. Surv. (CSUR)* **51**, 1–27 (2018).
71. Ito, N. et al. The accuracy and potential racial and ethnic biases of gpt-4 in the diagnosis and triage of health conditions: evaluation study. *JMIR Med. Educ.* **9**, e47532 (2023).
72. Benkirane, K., Kay, J. & Perez-Ortiz, M. How can we diagnose and treat bias in large language models for clinical decision-making? In *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (Volume 1: Long Papers) 2263–2288 (Association for Computational Linguistics, 2025).
73. Xiao, Y. et al. AMQA: an adversarial dataset for benchmarking bias of LLMs in medicine and healthcare. *arXiv* (2025).
74. Jin, R. et al. FairMedFM: fairness benchmarking for medical imaging foundation models. *Adv. Neural Inf. Process. Syst.* **37**, 111318–111357 (Curran Associates, Inc., 2024).
75. Fayyaz, H., Poulain, R. & Beheshti, R. Enabling scalable evaluation of bias patterns in medical LLMs. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2410.14763> (2024).
76. Goh, E. et al. ChatGPT influence on medical decision-making, bias, and equity: a randomized study of clinicians evaluating clinical vignettes. Preprint at *Medrxiv* <https://doi.org/10.1101/2023.11.24.23298844> (2023).
77. Swaminathan, A. et al. Feasibility of automatically detecting practice of race-based medicine by large language models. In *Proc. AAAI 2024 Spring Symposium on Clinical Foundation Models* (AAAI Press, 2024).
78. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *npj Digit. Med.* **6**, 195 (2023).
79. Chakraborty, J., Majumder, S. & Menzies, T. Bias in machine learning software: Why? How? What to do? In *Proc. 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* 429–440 (ACM, 2021).
80. Chakraborty, J., Majumder, S., Yu, Z. & Menzies, T. Fairway: A way to build fair ML software. In *Proc. 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* 654–665 (ACM, 2020).
81. Kamiran, F. & Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **33**, 1–33 (2012).
82. Chen, Z., Zhang, J., Sarro, F. & Harman, M. MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software. In *Proc. ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)* (ACM, 2022).
83. Peng, K., Chakraborty, J. & Menzies, T. Fairmask: better fairness via model-based rebalancing of protected attributes. *IEEE Trans. Softw. Eng.* **49**, 2426–2439 (2022).
84. Xiao, Y. et al. Mirrorfair: Fixing fairness bugs in machine learning software via counterfactual predictions. *Proc. ACM Softw. Eng.* **1**, 2121–2143 (2024).
85. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Volume 1: Long and Short Papers) 4171–4186 (Association for Computational Linguistics, 2019).
86. Lauscher, A., Lueken, T. & Glavaš, G. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021* 4782–4797 (Association for Computational Linguistics, 2021).
87. Liu, H. et al. Does gender matter? Towards fairness in dialogue systems. In *Proc. 28th International Conference on Computational Linguistics* 4403–4416 (International Committee on Computational Linguistics, 2020).
88. Huang, P.-S. et al. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020* 65–83 (Association for Computational Linguistics, 2020).
89. Gehman, S., Gururangan, S., Sap, M., Choi, Y. & Smith, N. A. RealToxicityPrompts: evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020* 3356–3369 (Association for Computational Linguistics, 2020).
90. Xu, J. et al. Recipes for safety in open-domain chatbots. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2010.07079> (2020).
91. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
92. Hasheminasab, S. A. et al. Assessing equitable use of large language models for clinical decision support in real-world settings: fine-tuning and internal-external validation using electronic health records from South Asia. Preprint at *medRxiv* <https://doi.org/10.1101/2024.06.05.24308365> (2024).
93. Yan, B. et al. The guideline for building fair multimodal medical AI with large vision-language model. Preprint at Research Square <https://doi.org/10.21203/rs.3.rs-5015239/v1> (2024).

94. Zahraei, P. S. & Shakeri, Z. Detecting bias and enhancing diagnostic accuracy in large language models for healthcare. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2410.06566> (2024).
95. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. In *Proc. ICLR* Vol. 1, 3 (OpenReview, 2022).
96. Zheng, G., Jacobs, M. A., Braverman, V. & Parekh, V. S. Towards fair medical AI: adversarial debiasing of 3D CT foundation embeddings. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2502.04386> (2025).
97. Munia, N. & Imran, A. A. Z. Prompting medical vision-language models to mitigate diagnosis bias by generating realistic dermoscopic images. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)* 1–4 (IEEE, 2025).
98. Khan, M. O., Afzal, M. M., Mirza, S. & Fang, Y. How fair are medical imaging foundation models? In *Proc. Machine Learning for Health (ML4H)* 217–231 (PMLR, 2023).
99. Hastings, J. Preventing harm from non-conscious bias in medical generative AI. *Lancet Digit. Health* **6**, e2–e3 (2024).
100. White, J. et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. In *Proc. 30th Conference on Pattern Languages of Programs (PLoP 2023)* (ACM, 2023).
101. Zhang, A., Yuksekogonul, M., Guild, J., Zou, J. & Wu, J. ChatGPT exhibits gender and racial biases in acute coronary syndrome management. Preprint at *medRxiv* <https://doi.org/10.1101/2023.11.14.23298525> (2023).
102. Ke, Y. H. et al. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. *J. Med. Internet Res.* **26**, e59439 (2024).
103. Schmidgall, S. et al. Addressing cognitive bias in medical language models. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2402.08113> (2024).
104. Wang, Y. et al. Unveiling and mitigating bias in mental health analysis with large language models. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2406.12033> (2024).
105. Hort, M., Chen, Z., Zhang, J. M., Harman, M. & Sarro, F. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM J. Responsible Comput.* **1**, 1–52 (2024).
106. Harrer, S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine* **90**, 104512 (2023).
107. Manès, V. J. et al. The art, science, and engineering of fuzzing: a survey. *IEEE Trans. Softw. Eng.* **47**, 2312–2331 (2019).
108. McKeeman, W. M. Differential testing for software. *Digit. Tech. J.* **10**, 100–107 (1998).
109. Singh, N., Lawrence, K., Richardson, S. & Mann, D. M. Centering health equity in large language model deployment. *PLOS Digit. Health* **2**, e0000367 (2023).
110. Chen, Z., Zhang, J. M., Sarro, F. & Harman, M. Fairness improvement with multiple protected attributes: How far are we? In *Proc. IEEE/ACM 46th International Conference on Software Engineering* 1–13 (ACM, 2024).
111. Chen, Z. et al. Software fairness dilemma: Is bias mitigation a zero-sum game? *Proc. ACM Softw. Eng.* **2**, 1780–1801 (2025).
112. Chen, Z., Li, X., Zhang, J. M., Sarro, F. & Liu, Y. Diversity drives fairness: Ensemble of higher order mutants for intersectional fairness of machine learning software. In *Proc. 47th IEEE/ACM International Conference on Software Engineering, ICSE 2025, Ottawa, ON, Canada, April 26 - May 6, 2025* 743–755 (IEEE, 2025).
113. Xiao, Y. et al. Supporting data for a scoping review of bias in large AI models for medicine and healthcare. <https://doi.org/10.5281/zenodo.20571512> (2026).
114. Junias, O., Kini, P. & Chaspari, T. Assessing algorithmic bias in language-based depression detection: a comparison of DNN and LLM approaches. In *Proc. 2025 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)* 1–7 (IEEE, 2025).
115. Ahmed, H. O. A. & Sadka, A. H. Auditing demographic bias in Mistral: an open-source LLM's diagnostic performance on the MedQA benchmark. *IEEE Access* **14**, 12526–12543 (2026).
116. Ji, Y., Zhang, H. & Wang, Y. Bias evaluation and mitigation in retrieval-augmented medical question-answering systems. In *Proc. AMIA Annual Symposium* Vol. 2025, 526 (AMIA, 2026).
117. Huang, A. S. et al. Chatgpt-assisted glaucoma diagnosis: a health-equitable multi-ancestry analysis using visual field and optical coherence tomography data. *Am. J. Ophthalmol.* **283**, 129–137 (2026).
118. Bai, N. et al. Detecting sociodemographic biases in the content and quality of large language model-generated nursing care: Cross-sectional simulation study. *J. Med. Internet Res.* **27**, e78132 (2025).
119. Xiao, X., Zhao, J., Payne, T. R. & Fang, M. Empirical study of social bias in medical question answering via large language models. In *Proc. International Conference on AI in Healthcare* 3–16 (Springer, 2025).
120. Wan, Z., Guo, Y., Bao, S., Wang, Q. & Malin, B. A. Evaluating sex and age biases in multimodal large language models for skin disease identification from dermoscopic images. *Health Data Sci.* **5**, 0256 (2025).
121. Chang, C. T. et al. Evaluating anti-lgbtqia+ medical bias in large language models. *PLOS Digit. Health* **4**, e0001001 (2025).
122. Balakrishnan, S. et al. Evaluating artificial intelligence bias in nephrology: the role of diversity, equity, and inclusion in AI-driven decision-making and ethical regulation. *Front. Artif. Intell.* **8**, 1525937 (2025).
123. Perez-Toro, P. A. et al. Exploring biases related to the use of large language models in a multilingual depression corpus: PaPerez-Toro et al. *Sci. Rep.* **15**, 36197 (2025).
124. Sun, X., Zou, X., Wu, Y., Wang, G. & Zhang, S. Fairness analysis of CLIP-based foundation models for X-ray image classification. In *Proc. 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)* 1–5 (IEEE, 2025).
125. Luo, Y. et al. FairDiffusion: enhancing equity in latent diffusion models via fair Bayesian perturbation. *Sci. Adv.* **11**, eads4593 (2025).
126. Jia, Q. et al. Implicit bias in safety-aligned large language models: a multi-faceted evaluation of clinical decision-making and health equity. *PLoS ONE* **21**, e0348819 (2026).
127. Fischetti, C. E. et al. Investigation of bias using ChatGPT-4 in pain management and patient profiling. *J. Emerg. Med.* **84**, 28–36 (2026).
128. Huang, Y. et al. Knowledge-guided adaptation of pathology foundation models effectively improves cross-domain generalization and demographic fairness. *Nat. Commun.* **16**, 11485 (2025).
129. Omar, M. et al. LLM-guided pain management: examining socio-demographic gaps in cancer vs non-cancer cases. Preprint at *medRxiv* <https://doi.org/10.1101/2025.03.04.25323396> (2025).
130. Nantakeeratipat, T. Large language model bias auditing for periodontal diagnosis using an ambiguity-probe methodology: a pilot study. *Front. Digit. Health* **7**, 1687820 (2025).
131. Ji, Y. et al. Mitigating the risk of health inequity exacerbated by large language models. *npj Digit. Med.* **8**, 246 (2025).
132. Andreadis, K. et al. Mixed methods assessment of the influence of demographics on medical advice of ChatGPT. *J. Am. Med. Inform. Assoc.* **31**, 2002–2009 (2024).
133. Goh, E. et al. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. *Commun. Med.* **5**, 59 (2025).
134. Liu, C. et al. Potential to perpetuate social biases in health care by Chinese large language models: a model evaluation study. *Int. J. Equity Health* **24**, 206 (2025).

135. Bouguettaya, A., Stuart, E. M. & Aboujaoude, E. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. *npj Digit. Med.* **8**, 332 (2025).
136. Sorin, V. et al. Socio-demographic modifiers shape large language models' ethical decisions. *J. Healthc. Inform. Res.* **9**, 567–586 (2025).
137. Levartovsky, A., Omar, M., Nadkarni, G. N., Kopylov, U. & Klang, E. Sociodemographic bias in large language model-assisted gastroenterology. *JAMA Netw. Open* **8**, e2532692 (2025).
138. Soffer, S. et al. Sociodemographic bias in large language model clinical trial screening. *J. Am. Med. Inform. Assoc.* **33**, 1504–1509 (2026).
139. Heinz, M. V. et al. Testing domain knowledge and risk of bias of a large-scale general artificial intelligence model in mental health. *Digit. Health* **9**, 20552076231170499 (2023).
140. Nijjer, K. et al. Adapting large language models to mitigate skin tone biases in clinical dermatology tasks: a mixed-methods study. Preprint at [arXiv https://doi.org/10.48550/arXiv.2510.00055](https://doi.org/10.48550/arXiv.2510.00055) (2025).
141. Sagers, L. W. et al. Augmenting medical image classifiers with synthetic data from latent diffusion models. Preprint at <https://doi.org/10.48550/arXiv.2308.12453> (2023).
142. Maslenskova, S. et al. Building trust in clinical LLMs: bias analysis and dataset transparency. In *Proc. 2025 Conference on Empirical Methods in Natural Language Processing 23021–23044* (Association for Computational Linguistics, 2025).
143. Zhang, Y. et al. CLIMB: a benchmark of clinical bias in large language models. Preprint at <https://doi.org/10.48550/arXiv.2407.05250> (2024).
144. Ghosh, A. et al. CLINIC: evaluating multilingual trustworthiness in language models for healthcare. Preprint at <https://doi.org/10.48550/arXiv.2512.11437> (2025).
145. Testoni, A. & Calixto, I. Calibrated? Not for everyone: how sexual orientation and religious markers distort LLM accuracy and confidence in medical QA. In *Proc. 64th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) 432–447 (Association for Computational Linguistics, 2026).
146. Rezaei, A. H. M. & Shakeri, Z. Counterfactual cultural cues reduce medical qa accuracy in LLMs: Identifier vs context effects. Preprint at <https://doi.org/10.48550/arXiv.2601.20102> (2026).
147. Rawat, R. et al. DiversityMedQA: a benchmark for assessing demographic biases in medical diagnosis using large language models. In *Proc. Third Workshop on NLP for Positive Impact* 334–348 (Association for Computational Linguistics, 2024).
148. Dai, S., Dai, W., Cheong, J. & Liang, P. P. FairGRPO: fair reinforcement learning for equitable clinical reasoning. (2025).
149. Xiao, Y. et al. FairMedQA: benchmarking bias in large language models for medical question answering. Preprint at <https://doi.org/10.48550/arXiv.2505.19562> (2026).
150. Gu, Z., Liu, Y., Zhang, Z. & Wang, S. Fairness-aware fine-tuning of vision-language models for medical glaucoma diagnosis. Preprint at <https://doi.org/10.48550/arXiv.2512.03477> (2026).
151. Li, D. et al. Fairness in multi-modal medical diagnosis with demonstration selection. Preprint at <https://doi.org/10.48550/arXiv.2511.15986> (2025).
152. Liu, M. et al. Gender bias in large language models for healthcare: assignment consistency and clinical implications. Preprint at <https://doi.org/10.48550/arXiv.2510.08614> (2025).
153. Zhang, Y., Dunn, A. G., Naseem, U. & Kim, J. Intersectional fairness in vision-language models for medical image disease classification. Preprint at <https://doi.org/10.48550/arXiv.2512.15249> (2026).
154. Haider, B., Gorti, A., Chadha, A. & Gaur, M. Mental health equity in LLMs: leveraging multi-hop question answering to detect amplified and silenced perspectives. In *Artificial Intelligence: Towards Sustainable Intelligence*, CCIS 2814, 42–59 (Springer, 2026).
155. Lamparth, M. et al. Moving beyond medical exams: a clinician-annotated fairness dataset of real-world tasks and ambiguity in mental healthcare. In *Proc. Fourteenth International Conference on Learning Representations (ICLR 2026)* (OpenReview, 2026).
156. Keough, P. PsychBench: auditing epidemiological fidelity in large language model mental health simulations. Preprint at <https://doi.org/10.48550/arXiv.2604.17359> (2026).
157. Liu, S., Zhang, S. & Bala, I. Robust or suggestible? Exploring non-clinical induction in LLM drug-safety decisions. Preprint at <https://doi.org/10.48550/arXiv.2510.13931> (2025).
158. Adiba, F. I. & Beheshti, R. Toward revealing nuanced biases in medical LLMs. Preprint at <https://doi.org/10.48550/arXiv.2507.21176> (2025).
159. Xia, Y. et al. Toward robust medical fairness: debiased dual-modal alignment via text-guided attribute-disentangled prompt learning for vision-language models. Preprint at <https://doi.org/10.48550/arXiv.2508.18886> (2025).
160. Koo, H. Towards robust and fair next visit diagnosis prediction under noisy clinical notes with large language models. Preprint at <https://doi.org/10.48550/arXiv.2511.18393> (2025).
161. Adappanavar, S., Shailya, K., Krishnan, G. S., Natarajan, S. & Ravindran, B. mFARM: towards multi-faceted fairness assessment based on harms in clinical decision support. Preprint at [arXiv https://doi.org/10.48550/arXiv.2509.02007](https://doi.org/10.48550/arXiv.2509.02007) (2025).
162. Ando, K. et al. A comparative study of English and Japanese ChatGPT responses to anaesthesia-related medical questions. *BJA Open* **10**, 100296 (2024).
163. Fiol, A. G. et al. Accuracy of Spanish and English-generated ChatGPT responses to commonly asked patient questions about labor epidurals: a survey-based study among bilingual obstetric anesthesia experts. *Int. J. Obstet. Anesth.* **61**, 104290 (2025).
164. Bilon, X. J. Can large language models support health literacy? Examining sociodemographic biases in ChatGPT's representation of hiv knowledge. *J. HIV Soc. Serv.* **24**, 8–28 (2026).
165. Wu, G. et al. Chatbots and diabetes: Is there gender bias? *J. Patient Exp.* **12**, 23743735251380954 (2025).
166. Wu, G., Paliath-Pathiyal, H., Khan, O. & Wang, M. C. Comparative analysis of LLMs in dry eye syndrome healthcare information. *Diagnostics* **15**, 1913 (2025).
167. Joshi, S. et al. Ensuring accuracy and equity in vaccination information from ChatGPT and CDC: mixed-methods cross-language evaluation. *JMIR Form. Res.* **8**, e60939 (2024).
168. Amin, K. S., Forman, H. P. & Davis, M. A. Even with ChatGPT, race matters. *Clin. Imaging* **109**, 110113 (2024).
169. Lozoya, D. C., D'Alfonso, S. & Conway, M. Identifying gender bias in generative models for mental health synthetic data. In *Proc. 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)* 619–626 (IEEE, 2023).
170. Reagen, C., Brown, K. M. & Hocker, T. Invisible bias in GPT-4o-mini: Detecting disparities in AI-generated patient messaging. *J. Med. Syst.* **49**, 127 (2025).
171. Lu, H. et al. Toward fair medical advice: addressing and mitigating bias in large language model-based healthcare applications. *Artif. Intell. Med.* **168**, 103216 (2025).
172. Schlicht, I. B., Zhao, Z., Sayin, B., Flek, L. & Rosso, P. Do LLMs provide consistent answers to health-related questions across languages? In *Advances in Information Retrieval: 47th European Conference on Information Retrieval (ECIR 2025)*, LNCS 15574, 314–322 (Springer, 2025).
173. Kondrup, E. & Imouza, A. Dr. bias: social disparities in AI-powered medical guidance. Preprint at <https://doi.org/10.48550/arXiv.2510.09162> (2025).
174. Xie, K. et al. Disparities in seizure outcomes revealed by large language models. *J. Am. Med. Inform. Assoc.* **31**, 1348–1355 (2024).

175. Rickman, S. Evaluating gender bias in large language models in long-term care. *BMC Med. Inform. Decis. Mak.* **25**, 274 (2025).
176. Huang, R. et al. Evaluation and bias analysis of large language models in generating synthetic electronic health records: comparative study. *J. Med. Internet Res.* **27**, e65317 (2025).
177. Nagesh, N. et al. FairTabGen: high-fidelity and fair synthetic health data generation from limited samples. Preprint at <https://doi.org/10.48550/arXiv.2508.11810> (2026).
178. Veenhuizen, M. & O'Malley, A. Demographic biases in AI-generated simulated patient cohorts: a comparative analysis against census benchmarks. *Adv. Simul.* **10**, 58 (2025).
179. Andrew, K. & Montalbano, M. J. Through a glass darkly: Perceptions of ethnoracial identity in artificial intelligence generated medical vignettes and images. *Med. Sci. Educ.* **35**, 1473–1488 (2025).
180. Qian, R. et al. Perturbation augmentation for fairer NLP. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing* 9496–9521 (Association for Computational Linguistics, 2022).
181. Lu, K., Mardziel, P., Wu, F., Amancharla, P. & Datta, A. Gender bias in neural natural language processing. In *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday* (eds Nigam, V. et al.) Lecture Notes in Computer Science 12300, 189–202 (Springer, Cham, 2020).
182. Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M. & Goldberg, Y. Null it out: guarding protected attributes by iterative nullspace projection. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 7237–7256 (Association for Computational Linguistics, 2020).
183. Iskander, S., Radinsky, K. & Belinkov, Y. Shielded representations: protecting sensitive attributes through iterative gradient-based projection. In *Findings of the Association for Computational Linguistics: ACL 2023* 5961–5977 (Association for Computational Linguistics, 2023).
184. Bartl, M., Nissim, M. & Gatt, A. Unmasking contextual stereotypes: measuring and mitigating BERT's gender bias. In *Proc. Second Workshop on Gender Bias in Natural Language Processing* 1–16 (Association for Computational Linguistics, 2020).
185. Han, X., Baldwin, T. & Cohn, T. Balancing out bias: achieving fairness through balanced training. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing* 11335–11350 (Association for Computational Linguistics, 2022).
186. Tokpo, E. K. & Calders, T. Text style transfer for bias mitigation using masked language modeling. In *Proc. 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop* 163–171 (Association for Computational Linguistics, 2022).
187. Dhingra, H., Jayashanker, P., Moghe, S. & Strubell, E. Queer people are people first: deconstructing sexual identity stereotypes in large language models. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2307.00101> (2023).
188. Yang, K., Yu, C., Fung, Y. R., Li, M. & Ji, H. Adept: A debiasing prompt framework. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 37, 10780–10788 (AAAI Press, 2023).
189. Fatemi, Z., Xing, C., Liu, W. & Xiong, C. Improving gender fairness of pre-trained language models without catastrophic forgetting. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers) 1249–1262 (Association for Computational Linguistics, 2023).

Author contributions

Y.X., Z.C., and J.M.Z. conceived the study and developed the methodology. Y.X. conducted the literature search and screening. Y.X. and Z.C. performed data extraction and analysis. Y.X. wrote the initial draft. J.H., W.C., Y.L., K.L., M.R.M., R.D., and J.M.Z. reviewed and edited the manuscript. Z.C. and J.M.Z. supervised the project. All authors approved the final manuscript.

Funding

Z.C. discloses support for the research of this work from the New Generation Artificial Intelligence-National Science and Technology Major Project (grant no. 2025ZD0123803). J.M.Z. discloses support for the research of this work from ITEA grants GreenCode (project no. 23016) and GENIUS (project no. 23026). The other authors declare no relevant funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44528-026-00041-2>.

Correspondence and requests for materials should be addressed to Zhenpeng Chen.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026