Bias in Large AI Models for Medicine and Healthcare: Survey and Challenges

Ying Xiao[1], Zhenpeng Chen[2†], Jen-tse Huang[3], Wenting Chen[4], Yepang Liu[5], Kezhi Li[6], Mohammad Reza Mousavi[1], Richard Dobson[1] and Jie M. Zhang[1]

[1]King's College London, [2]Nanyang Technological University, [3]Johns Hopkins University, [4]Stanford University, [5]Southern University of Science and Technology, [6]University College London

Large AI models have demonstrated human-expert-level performance in specific medical domains. However, concerns regarding medical bias have prompted growing attention from the medicine, sociology, and computer science communities. Although research on medical bias in large AI models is rapidly expanding, efforts remain fragmented, often shaped by discipline-specific assumptions, terminology, and evaluation criteria. This survey provides a comprehensive synthesis of 55 representative studies, organizing the literature into three core themes: taxonomy of medical bias, methods for detection, and strategies for mitigation. Our analysis bridges the conceptual and methodological gaps across disciplines and highlights persistent challenges, including the lack of unified foundations for medical fairness, insufficient datasets and evaluation benchmarks, the lack of methods for rigorous automatic bias detection, missing real-world validation and continuous validation, inadequate representation, as well as insufficient studies on the trade-off between fairness and accuracy. Thereby, we identify and highlight emerging research opportunities to address these gaps. To further advance the field, we present a structured index of publicly available large AI models and datasets referenced in these studies.

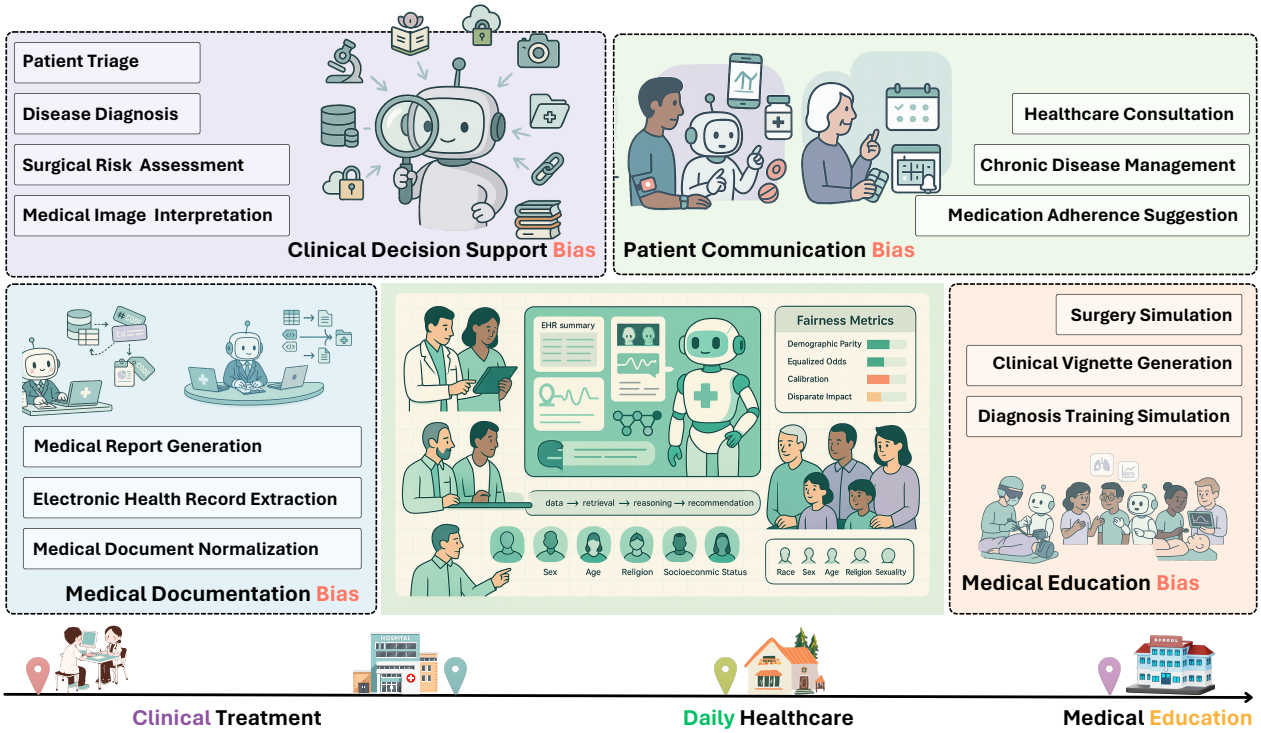

Figure 1 | An overview of bias in Large AI models for medicine and healthcare.

Corresponding author: †Zhenpeng Chen. Email: zhenpeng.chen@ntu.edu.sg

Contents

1. Introduction

Artificial Intelligence (AI) is increasingly integral to modern healthcare. Among these technologies, large AI models, including large language models (LLMs) and large vision models (LVMs), have recently emerged as especially influential (Bommasani et al., 2021). Trained on massive corpora, these advanced models, exemplified by ChatGPT, demonstrate human-like understanding and generation across multiple data modalities (Achiam et al., 2023; Bommasani et al., 2021). Large, general-purpose AI models are now being deployed or studied across a spectrum of healthcare applications (D'Antonoli et al., 2024; Gao et al., 2023; He et al., 2024; Kim et al., 2024; McDuff et al., 2023), including but not limited to:

- Clinical decision support – Assisting the clinicians with clinical diagnostic reasoning or treatment planning based on large knowledge bases and patient-specific data (Chansiri et al., 2024; Poulain et al., 2024a; Yang et al., 2025b).
- Patient communication – Powering conversational agents or chatbots that answer patient questions, provide triage advice, or offer health counseling in natural language (Nastasi et al., 2023; Pfohl et al., 2024; Poulain et al., 2024b).
- Medical documentation – Extracting, summarizing, interpreting, or generating clinical notes and medical reports, therefore reducing the documentation burden of clinicians (Chansiri et al., 2024; Hanna et al., 2023, 2025).
- Medical education – Producing understandable medical explanations, personalized education materials, or providing specific training simulation for medical professionals (Agrawal, 2024; Chansiri et al., 2024; Chen et al., 2024a).

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 (Nazi and Peng, 2024; Omiye et al., 2024; Pahune and Rewatkar, 2023).

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 (Gallegos et al., 2024; Kim et al., 2025). In this survey, we define bias as **any systematic error, stereotype, or prejudice in an AI system's outputs that disadvantages certain individuals or groups**, thereby undermining fairness. A wide range of bias issues in large AI models for health have been reported. For instance, Czum and Parr (2023) 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 (Bommasani et al., 2021; Singhal et al., 2023a). Similar demographic gaps have been documented across racial and socioeconomic groups, further threatening trust in AI-powered healthcare (Apakama et al., 2024; Kim et al., 2023; Nastasi et al., 2023; Pfohl et al., 2024; Yeo et al., 2025).

While these biases are deeply concerning, it is important to recognize that they often originate from the limitations and biases inherent in their training data, reflecting historical and societal inequities captured in healthcare records and research (Guo et al., 2024; Jones et al., 2024; Kim et al., 2025; Li et al., 2023a; Pfohl et al., 2024). Nevertheless, large AI models, despite inheriting such biases from their training data, also offer a unique opportunity: the systematic identification and mitigation of these biases at scale, using various techniques such as training data rebalancing, adversarial debiasing, and output adjustment. The solutions are often more feasible and scalable than altering deeply ingrained, complex human biases. This creates an unprecedented opportunity

for human-AI partnership, ultimately achieving 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—rising from 15 in 2023 to 55 by June 2025—conceptual fragmentation and terminological inconsistencies remain common (Chen et al., 2024a; Pfohl et al., 2024). These efforts can be fragmented, each 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 (Chen et al., 2024a,b; Luo et al., 2024a; Tian et al., 2023). Clinical researchers focus on the damage, risk, and ethics 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 (Nastasi et al., 2023; Pfohl et al., 2024). 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 common foundation and facilitate collaborative progress toward fair and trustworthy AI-driven healthcare, we present the novel survey that systematically examines medical bias in large AI models, aiming to unify theoretical frameworks, spotlight methodological innovations, and identify shared challenges across communities. In particular, we present the application of large AI models in medicine and healthcare, introduce the concept of medical bias in these models, describe the types and distribution of medical bias, review current technologies for detecting and mitigating medical bias, and conclude with a discussion of existing challenges and potential research opportunities.

To summarize, our contributions include:

- A detailed conceptual framework for medical bias in large AI models, synthesizing perspectives from artificial intelligence, clinical medicine, and healthcare policy;
- A comprehensive and up-to-date synthesis of 55 representative studies, categorizing detection and mitigation strategies by technical approach and clinical scenario;
- An in-depth analysis of persistent challenges in addressing medical bias and a discussion of promising research opportunities in achieving fairer AI medicine and healthcare;
- A curated index of publicly available large AI models and datasets referenced in the surveyed literature, enabling easier access and reproducibility for future research.

This survey is structured as follows. Section 2 presents the preliminaries, including the definition and evaluation of medical bias. Figure 2 presents a roadmap of bias in large AI models for medicine and healthcare, covering the taxonomy of medical bias along with the detection and mitigation strategies. In Section 3, we describe our survey methodology. We then review existing work, categorizing it into bias by medical scenarios and clinical domains (Section 4), bias detection (Section 5), and bias mitigation (Section 6). Next, we introduce the large AI models and datasets associated with the collected publications (Section 7) and analyze research trends and distributions (Section 8.1). We then discuss open problems, research opportunities, and threats to validity in Section 8.2. Finally, we conclude the survey in Section 9.

2. Preliminaries

In this section, we introduce the background knowledge of large AI models for health, followed by a conceptual overview of bias and its potential causes in large AI models.

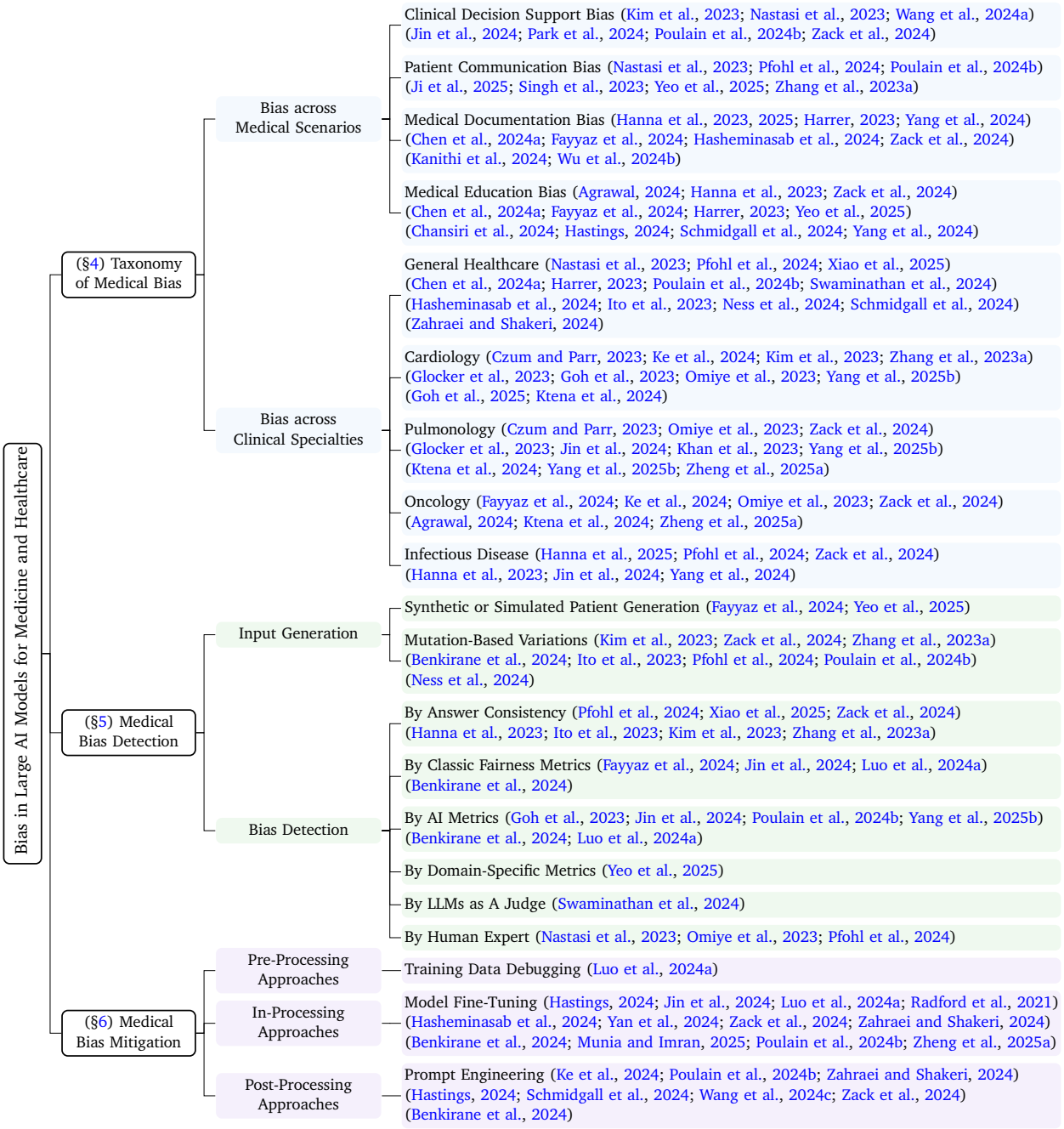

Figure 2 | A taxonomy of medical biases along with their corresponding detection and mitigation approaches.

2.1. Background of Large AI Models

Large AI models, often referred to as *Foundation Models*, are massive deep learning models (typically based on Transformer architectures) that are first pre-trained on broad data at an immense scale, usually via self-supervised learning, and then adapted or fine-tuned for specific tasks (Zhao et al., 2023a). The rise of large AI models has been enabled by three key factors: (1) unprecedented volumes of training data, (2) the Transformer architecture (Vaswani et al., 2017) and other scalable neural network designs, and (3) vast increases in compute power for training at scale (Zhao et al., 2023a). Crucially, these models can be fine-tuned or prompted to perform a wide range of downstream tasks,

making them highly versatile compared to traditional task-specific AI models.

The main categories of large AI models are LLMs, LVMs, and large multimodal models (LMMs). In the following, we briefly outline what they are, how they are trained, and why they have become so prominent in recent years.

2.1.1. Large Language Models

LLMs are large AI models that specialize in human language. The training process usually involves a self-supervised objective such as autoregressive next-word prediction or masked language modeling, so that the model learns linguistic patterns, grammar, facts, and reasoning abilities without needing manually labeled examples (Zhao et al., 2023a). The Transformer's self-attention mechanism enables learning long-range dependencies in text efficiently, which is also a key breakthrough leading to modern LLMs (Zhao et al., 2023a).

In 2018, OpenAI released the first Generative Pre-trained Transformer (GPT) model, introducing a new approach to language modeling based on unsupervised pretraining and fine-tuning (Radford et al., 2018). This was followed by successive iterations: GPT-2 (2019) (Radford et al., 2019; Solaiman et al., 2019), GPT-3 (2020) (Brown et al., 2020), GPT-3.5 (2022) (OpenAI, 2022), GPT-4 (2023) (Achiam et al., 2023), and GPT-5 (2025) (OpenAI, 2025). With GPT-3 and GPT-3.5, many observers noted a significant step change in generative performance (Kalyan, 2024; Ye et al., 2023), with models demonstrating fluency, coherence, and task generalization that surpassed earlier systems. This leap attracted widespread attention to GPT and ChatGPT in particular, and to LLMs more broadly. In addition, leading AI companies such as Meta, Anthropic, xAI, Mistral AI, Alibaba, DeepSeek, and Moonshot have each launched a series of powerful large language models, including Llama-4 (2025) (AI, 2025a), Claude-4 (2025) (Anthropic, 2025a), Grok-4 (2025) (xAI, 2025), Mistral Medium 3.1 (2025) (MistralAI, 2025), Qwen3 (2025) (Yang et al., 2025a), DeepSeek-V3.1 (2025) (DeepSeek, 2025), and Kimi K2 (2025) (Kimi et al., 2025), respectively.

2.1.2. Large Vision Models

Large vision models (LVMs) refer to high-capacity models trained on massive image datasets. A key milestone was the Vision Transformer (ViT) (Dosovitskiy et al., 2020), which applied the Transformer architecture to images and demonstrated state-of-the-art performance when trained on large-scale datasets. ViT treated image patches as tokens, analogous to words in text, and showed that scaling model size and data leads to broadly useful visual representations.

Another major breakthrough was OpenAI's CLIP model (Radford et al., 2021), trained on 400 million image–text pairs to learn a joint embedding space for vision and language. CLIP enabled zero-shot image recognition using natural language prompts, demonstrating the potential of language-supervised vision models. This was followed by diffusion-based generative models (e.g., DALL·E 2 (Ramesh et al., 2021)), which generate high-quality images from text inputs.

2.1.3. Large Multimodal Models

Large multimodal models (LMMs), also known as Multimodal LLMs, process multiple data types, such as text, images, audio, or structured data, within a single system. They aim to integrate diverse modalities to understand complex, real-world inputs and generate rich outputs. An influential example is DeepMind's Flamingo (Alayrac et al., 2022), which links vision and language components to perform few-shot multimodal learning. OpenAI's GPT-4 (Achiam et al., 2023) also accepts image inputs, enabling visual question answering and diagram interpretation. These models are built by

combining LLMs with vision encoders and training them on multimodal datasets. LMMs are rapidly advancing the field toward more general, flexible, and human-like AI.

2.2. Development of Large AI Models for Medicine and Healthcare

The past few years have seen a dramatic shift in the scale and capabilities of AI models applied to healthcare. Early successes of deep learning in medicine came from task-specific models, such as convolutional neural networks for medical image analysis (e.g., detecting pneumonia on chest X-rays) (Salehi et al., 2023) and recurrent models for clinical text (Liu et al., 2017). However, these models required large labeled datasets and were limited to narrow tasks. The introduction of transformer architectures and pre-trained language models revolutionized this landscape, enabling large AI models that learn from massive unlabeled corpora and can be adapted to various downstream healthcare tasks. BioBERT (Lee et al., 2020) was a landmark model that fine-tuned Google's BERT (Devlin et al., 2019) on 18 billion words of biomedical literature, significantly improving biomedical named entity recognition and question answering tasks over general BERT. Likewise, ClinicalBERT was trained on clinical notes to better understand healthcare narratives (Huang et al., 2019a). These early medical transformers laid the groundwork but were relatively small in scale and lacked generative abilities for text generation tasks.

Inspired by the success of GPT-3 in general domains, the community began developing larger LLMs tailored to medical data. A notable advance was BioGPT (Luo et al., 2022) by Microsoft, a generative Transformer with ~1.5 billion parameters trained on 15 million medical research abstracts from a database of the U.S. National Library of Medicine. BioGPT was among the first domain-specific GPT-style models, introducing text generation capabilities (e.g., generating biomedical hypotheses or summaries) beyond the scope of earlier BERT models. It achieved state-of-the-art results on biomedical question answering and information extraction tasks, demonstrating the power of large AI models in biomedical text mining (Luo et al., 2022).

2.3. Bias in Large AI Models

The history of AI bias traces back to the earliest applications of machine learning, where models trained on historical data began to reflect and perpetuate societal inequalities embedded in those datasets (Hort et al., 2024). As AI models spread to areas such as criminal justice, finance, and healthcare, researchers observed that they frequently produced discriminatory outcomes, favoring majority populations and disadvantaging underrepresented groups. For instance, in healthcare, concerns deepened when commercial risk-scoring algorithms were shown to underestimate the health needs of black patients due to flawed proxies such as historical healthcare spending (Obermeyer et al., 2019). These revelations spurred the emergence of AI fairness as a formal area of study, prompting the development of fairness metrics, bias mitigation strategies, and policy guidance aimed at ensuring AI supports equitable outcomes (Huang et al., 2025b).

2.3.1. Classic AI Bias and Healthcare AI Bias

In classical AI research, bias is typically defined through the lens of statistical fairness, where the goal is to ensure that predictive outcomes are equal across different demographic groups and that sensitive attributes (also called protected attributes) such as race, gender, or age do not unjustly influence model decisions (Du et al., 2025; Huang et al., 2025a; Shi et al., 2025; Wang et al., 2024b). As a result, sensitive attributes are often excluded from models to prevent discriminatory outcomes. In healthcare AI, however, this notion of fairness becomes more nuanced and domain-specific. **Sensitive attributes may carry clinically relevant information**. For example, race can correlate with genetic

risk factors or social determinants of health, and sex differences may influence disease presentation or treatment response (Rajkomar et al., 2018). Thus, in medical contexts, excluding such attributes may actually significantly harm predictive accuracy or contribute to poorer health outcomes for certain groups. This tension has made the healthcare AI bias detection as well as mitigation much more challenging.

Furthermore, the identification of sensitive attributes is context-dependent and varies across medical scenarios and diseases. Based on a review of existing literature, commonly studied sensitive attributes include race, gender, age, ethnicity, disability status, religious beliefs, socioeconomic status, language, and geographic location, among others (Gallegos et al., 2024; Guo et al., 2024; Pfohl et al., 2024; Poulain et al., 2024b).

2.3.2. Causes of Healthcare AI Bias

Unfairness in large AI models for health can emerge at multiple stages of the model life cycle (Harrer, 2023).

Biased labels in training data: A major source of unfairness in large AI models is bias embedded in the labels of training data. Since these models are often trained on massive datasets collected from sources such as the internet or medical records, they inevitably inherit societal inequalities reflected in the labels. Consequently, LLMs can absorb and reproduce these biases in their outputs (Li et al., 2025; Shahbazi et al., 2023).

Underrepresentation of minority groups: If certain racial, ethnic, or other demographic groups are sparsely represented in the training corpus, the model's performance on inputs concerning those groups will be affected (Mehrabi et al., 2021).

Linguistic and cultural variation: Differences in language use and cultural context further contribute to unfairness. Most LLMs are predominantly trained on English-language sources and standard writing styles, meaning they may struggle with non-standard dialects, multilingual inputs, or culturally specific terminology. This linguistic and cultural variation issue can cause the model to misinterpret or inadequately respond to patients who use vernacular speech, idioms from different cultures, or languages other than English (Tierney et al., 2025).

Model evaluation and optimization: Relying primarily on aggregate performance metrics or on majority populations during model selection and hyperparameter tuning may obscure group-specific performance gaps, thereby reinforcing hidden disparities (Guo et al., 2024).

Model alignment: A further challenge is introduced during the alignment phase, where models are fine-tuned using human feedback or optimization techniques to encourage desirable behavior. If alignment is conducted with annotators or guidelines that do not reflect diverse clinical expertise, cultural sensitivity, or ethical values, the resulting model may systematically favor normative or majority viewpoints (Bai et al., 2022).

Model deployment: Bias can also be caused by mismatch between the training environment and real-world use cases, or pragmatic use by end users (e.g., patients, clinicians, and healthcare systems) in real-world settings (Singh et al., 2023). For instance, when LLMs trained on data from urban hospitals are applied to rural or underserved populations, leading to less accurate or biased recommendations for minority groups. The lack of transparency and interpretability in LLMs can make it difficult for healthcare professionals and patients to identify and mitigate bias post hoc, further compounding disparities in care (Gallegos et al., 2024; Singh et al., 2023).

3. Survey Methodology

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

3.1. Survey Scope

This survey investigates the emerging interdisciplinary field of *bias in large AI models for medicine and healthcare*. As illustrated in Figure 3, the scope lies at the intersection of medicine and healthcare, bias and fairness, and large AI model research. Our goal is to establish a coherent understanding of how large AI models exhibit, detect, or mitigate bias in healthcare contexts by systematically organizing conceptual definitions, empirical findings, and technical contributions. We include papers that meet at least one of the following criteria: (1) define or characterize medical bias in the context of large AI models; (2) propose methods, frameworks, or tools for detecting or measuring medical bias in large AI models; or (3) present strategies for mitigating bias in medicine and healthcare tasks involving large AI models.

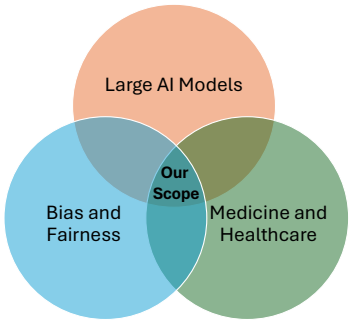

Figure 3 | The scope of our survey is located at the intersection of the three fields.

Building on our inclusion criteria, we exclude studies that (1) focus exclusively on traditional machine learning models (e.g., logistic regression and support vector machines) or simple deep neural networks without engaging large AI models; (2) discuss general fairness principles or bias outside the medical and healthcare domains; or (3) examine diversity, inclusion, or cognitive bias in clinical workflows without analyzing model-generated outputs. This survey centers on the medical bias of large AI models in medicine and healthcare as reflected in system outputs and decision behaviors, rather than on institutional or societal disparities independent of model behaviors.

3.2. Paper Collection

To construct a comprehensive corpus of relevant literature, we performed keyword-based searches in five major academic databases: ACM Digital Library, IEEE Xplore, Web of Science, PubMed, and Google Scholar. 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 (Liu et al., 2024b; Zheng et al., 2025b). We also included publications from open-access preprint servers, i.e., arXiv and medRxiv. This selection strategy ensures access to both peer-reviewed and emerging research across disciplines.

The search string was developed through an iterative trial-and-error process (Chen et al., 2024b; Lin et al., 2022). We started with broad queries— *("bias" AND "medical" AND "large AI model")* and *("bias" AND "medical" AND "foundation model")*—and then 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 helped improve coverage and ensure relevance to the topic of medical bias in large AI models (i.e., foundation models).

The final search string used was: *("bias" OR "fair" OR "discrimination" OR "equity") AND ("large AI model" OR "foundation model" OR "multimodal model" OR "vision model" OR "language model" OR "GPT" OR "LLM" OR "LVM" OR "LMM") AND ("medical" OR "medicine" OR "diagnosis" OR "health" OR "clinic" OR "surgery" OR "patient" OR "treatment")*.

We executed separate searches across the five selected academic databases on June 17, 2025. This yielded 328 results from IEEE Xplore, 335 from PubMed, 12 from the ACM Digital Library, 293 from Web of Science, and approximately 258,100 from Google Scholar. For Google Scholar, we screened only the top 300 results ranked by relevance. The authors manually reviewed the title and abstract of each paper to determine its alignment with our inclusion criteria, yielding 42 relevant papers. To further enhance completeness and reduce potential selection bias, we employed backward and forward snowballing (Jalali and Wohlin, 2012). In backward snowballing, we examined the reference lists of included papers. In forward snowballing, we used Google Scholar to identify citing articles. This procedure was repeated until no additional relevant works were found. It yielded 13 additional papers. In total, we included 55 papers in this survey.

3.3. Paper Analysis

To systematically analyze the selected body of literature, we conducted a thematic synthesis following established guidelines for qualitative review methods (Huang et al., 2018). This approach enables structured organization and synthesis of findings across diverse studies, particularly suitable for emerging interdisciplinary domains such as medical bias in large AI models.

The first two authors manually reviewed the full text of each included paper. During this process, we extracted structured information regarding (1) the definition and conceptualization of medical bias, (2) bias detection methodologies, (3) mitigation strategies, and (4) associated medical domains, demographic attributes, datasets, and evaluation metrics. Through iterative grouping, we distilled three high-level thematic categories: (1) conceptual foundations of medical bias, (2) techniques for detecting medical bias, and (3) strategies for mitigating bias in large AI models for medicine and healthcare tasks. Finally, all authors independently double-checked the content, reviewing it for potential errors, inconsistencies, or omissions.

4. Taxonomy of Medical Bias in Large AI Models

Medical bias in large AI models manifests in diverse forms across clinical settings and disease domains. To systematically characterize these variations, we develop a dual taxonomy that organizes the literature along two principal axes: (1) medical scenarios, such as decision support and education; and (2) clinical specialties, such as Cardiology and Pulmonology. This dual taxonomy facilitates nuanced analysis of medical bias from both functional and clinical perspectives, delineating the wide range of contexts in which bias emerges. By structuring the literature in this way, we enable more precise identification of high-risk scenarios and specialties, laying the groundwork for targeted evaluation and mitigation strategies.

Table 1 | Examples of medical bias across scenarios.

Medical Bias Type	Example	Example Explanation
Clinical Decision Support Bias (Akhras et al., 2025; Benkirane et al., 2024; Chansiri et al., 2024; Czum and Parr, 2023; Fayyaz et al., 2024; Glocker et al., 2023; Goh et al., 2023, 2025; Harrer, 2023; Hasheminasab et al., 2024; Hastings, 2024; Ito et al., 2023; Jin et al., 2024; Kanithi et al., 2024; Khan et al., 2023; Kim et al., 2023; Ktena et al., 2024; Luo et al., 2024a; Munia and Imran, 2025; Ness et al., 2024; Omar et al., 2025; Omiye et al., 2023; Pfohl et al., 2024; Poulain et al., 2024a; Rani et al., 2024; Schmidgall et al., 2024; Schnepper et al., 2025; Singh et al., 2023; Swaminathan et al., 2024; Templin et al., 2025; Wang et al., 2024c; Wu et al., 2024b; Xiao et al., 2025; Yan et al., 2024; Yang et al., 2024, 2025b; Young et al., 2022; Zack et al., 2024; Zahraei and Shakeri, 2024; Zheng et al., 2025a)	Researchers provided clinical vignettes related to coronary artery disease, manually altering the gender 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., 2023).	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 gender 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 (Ji et al., 2025; Nastasi et al., 2023; Pfohl et al., 2024; Poulain et al., 2024b; Singh et al., 2023; Yeo et al., 2025; Zhang et al., 2023a)	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] [Gender] 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., 2023)	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 (Chen et al., 2024a; Fayyaz et al., 2024; Hanna et al., 2023, 2025; Harrer, 2023; Hasheminasab et al., 2024; Kanithi et al., 2024; Wu et al., 2024b; Yang et al., 2024; Zack et al., 2024)	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., 2024).	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 (Agrawal, 2024; Chansiri et al., 2024; Chen et al., 2024a; Fayyaz et al., 2024; Hanna et al., 2023; Harrer, 2023; Hastings, 2024; Schmidgall et al., 2024; Yang et al., 2024; Yeo et al., 2025; Zack et al., 2024)	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., 2024).	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.

4.1. 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 1, 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, as observed by (Omiye et al., 2023).

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, provide triage advice, 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 (Akhras et al., 2025; Kanithi et al., 2024; Schnepper et al., 2025; Yeo et al., 2025; Zack et al., 2024).

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 (Hanna et al., 2023; Yang et al., 2024). 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.

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 assumptions (Agrawal, 2024; Zack et al., 2024). Such imbalances can distort both downstream models and medical education materials, misrepresenting the real-world diversity of diseases and populations.

4.2. 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 Cardiology. Large AI models frequently show demographic bias in cardiology diagnosis and treatment recommendations. For example, it was reported that GPT-4 altered its diagnosis when only gender changed from black male to black female, a discrepancy not seen in clinicians (Kim et al., 2023). Significant disparities in angiography recommendation and false negative rates for black females are also reported (Yang et al., 2025b). Such patterns likely arise from data imbalance and latent stereotypes, risking unequal cardiac care.

Medical Bias in Pulmonology. In pulmonology, large AI models consistently show lower diagnostic performance for female and Black patients. Large-scale studies report reduced Area Under the Curve (AUC) values for these groups in chest disease diagnosis (Czum and Parr, 2023; Glocker et al., 2023; Khan et al., 2023). Bias also observed in AI-generated clinical vignettes, where certain ethnicities are disproportionately represented (Zack et al., 2024). These disparities likely stem from imbalanced training data and optimization objectives that prioritize overall accuracy over subgroup equity.

Medical Bias in Infectious Disease. Large AI models show substantial bias in diagnostic output and prevalence estimation of infectious diseases. Existing studies (Hanna et al., 2023; Yang et al., 2024; Zack et al., 2024) found that GPT-4 and GPT-3.5 displayed significant race and gender disparities in HIV/COVID reports and diagnoses. Such bias can propagate or amplify real-world health inequities.

Medical Bias in Oncology. Large AI models frequently misestimate cancer prevalence across demographic groups, with reported gaps exceeding 40 percentage points for certain cancers in black or male populations (Zack et al., 2024). These errors likely stem from training on incomplete or biased epidemiological data.

Medical Bias in Dermatology. Large AI models for dermatology often underperform on minority skin tones and exhibit gender treatment bias. For instance, male patients were more likely to be prescribed isotretinoin for acne, and fairness gaps in skin cancer detection can be reduced by demographic-aware modeling (Kim et al., 2023; Ktena et al., 2024).

Medical Bias in Rheumatology and Immunology. In rheumatology and immunology, biases span

diagnosis, prevalence estimation, and vignette generation. For example, GPT-4 overrepresented female RA patients in generated cases and showed prevalence estimation gaps exceeding 30 percentage points (Kim et al., 2023; Yang et al., 2024; Zack et al., 2024). These findings reflect both data-driven and algorithmic sources of unfairness.

Bias in Psychiatry. Yeo et al. (2025) studied whether GPT-4 had sociodemographic bias in mental health support. Their study did not uncover significant evidence of bias within an LLM-enabled mental health conversational agent.

Medical Bias in Ophthalmology. Large AI models in ophthalmology exhibit race- and gender-related bias, primarily driven by imbalanced ocular imaging datasets. Demographic parity and equalized odds differences in glaucoma diagnosis are notably high with standard Contrastive Language-Image Pre-Training (CLIP) models, but can be reduced through targeted mitigation (Luo et al., 2024a; Radford et al., 2021; Yan et al., 2024).

Across clinical domains, race, gender, 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.

5. Medical Bias Detection

Bias detection (also known as bias testing and fairness testing) aims to identify and quantify potential biases and unfairness in AI models(Chen et al., 2024b; Zhang and Harman, 2021). Existing studies employ two major types of criteria for bias detection: answer consistency and statistical measurements. Additionally, in some complex healthcare scenarios, where answer consistency and statistical assessment may be unavailable, human-based bias assessments provide a flexible alternative. In the following, we introduce these types of bias assessment criteria in detail.

5.1. 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 of large AI models' output (Pfohl et al., 2024). 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 (Chen et al., 2018; Segura et al., 2016). These strategies can be broadly categorized into synthetic or simulated patient generation and mutation-based variations of existing cases.

5.1.1. Synthetic or Simulated Patient Generation

Yeo et al. (2025) utilized simulated patients derived from digital standardized patients (DSPs) to assess biases in GPT-4's provision of mental health support. Fayyaz et al. (2024) introduced a vignette generation method for scalable, evidence-based bias evaluation in medical LLMs. Their procedure incorporated domain-specific bias characterization, mitigation of hallucinations, and dependencies between health outcomes and sensitive attributes, leveraging medical knowledge

graphs and ontologies. They applied this to case studies on obesity, breast cancer, prostate cancer, and pregnancy, demonstrating that the generated test cases reliably uncover bias patterns in LLMs.

5.1.2. Mutation-Based Variations

Many studies mutate existing prompts or cases by systematically altering sensitive attributes such as gender, race, or ethnicity.

Several studies employed manual or structured attribute variations. Zack et al. (2024) adapted 19 expert-generated medical education cases from the NEJM Healer platform by varying patient gender (male or female) and race/ethnicity (Asian, Black, White, or Hispanic). Zhang et al. (2023a) permuted race (none, Caucasian, African American, Hispanic) and gender (none, female, male) in prompts to assess biases in GPT-3.5 for acute coronary syndrome (ACS) management. Kim et al. (2023) constructed 19 clinical vignettes spanning multiple specialties, systematically varying gender, race/ethnicity, 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. (2023) 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 race or ethnicity. Benkirane et al. (2024) 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 gender-neutral versions along with ethnic variations (Arab, Asian, Black, Hispanic, White). Pfohl et al. (2024) 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. (2024b) applied red-teaming strategies, mutating questions via adversarial prompting and rotating through different patient demographics. Ness et al. (2024) proposed *MedFuzz*, which uses an LLM as an attacker to automatically generate adversarial inputs that elicit biased behavior in medical scenarios. For example, in a case of a 6-year-old African American boy with anemia and jaundice, GPT-4 correctly diagnosed sickle cell disease; however, after adding descriptors such as low-income status, herbal remedy use, and immigrant parents, MedFuzz induced GPT-4 to incorrectly diagnose hemoglobin C disease.

5.2. Bias Evaluation

5.2.1. 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, gender, socioeconomic status) differ. Pfohl et al. (2024) 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. (2025) 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. Hanna et al. (2023); Ito et al. (2023); Kim et al. (2023); Zack et al. (2024); Zhang et al. (2023a) 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).

5.2.2. 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. (2024) applied Demographic Parity, Equal Opportunity, and Equalized Odds in their FairMedFM benchmark for medical imaging. Luo et al. (2024a) introduced Harvard-FairVLMed, a vision–language medical dataset with demographic attributes, ground-truth labels, and clinical notes, enabling fairness analysis in large vision models; they reported results using Demographic Parity, Equalized Odds, and AUC differences. Fayyaz et al. (2024) employed Demographic Parity and Equal Opportunity to measure fairness in their evaluations. Benkirane et al. (2024) 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.

5.2.3. Bias Detection with AI Metrics

Beyond classic fairness measures, many studies quantify the bias in large health AI models using AI performance metrics tailored to specific tasks. Yang et al. (2025b) measured underdiagnosis disparity in chest X-ray interpretation by comparing false-negative rates (FNR) and false-positive rates (FPR) across race, sex, and age groups. Jin et al. (2024) evaluated disparities in accuracy, AUC, and Dice Similarity Coefficient (DSC) in their medical imaging benchmark, and—alongside Poulain et al. (2024b)—compared predictive probability distributions across demographic groups, applying statistical tests to assess significance. Benkirane et al. (2024) used accuracy consistency and SHAP value analysis to quantify bias in LLM-driven clinical decision-making. Goh et al. (2023) compared clinical decision accuracy between White men and Black women in chest pain evaluation. Luo et al. (2024a) measured fairness in glaucoma diagnosis from a vision–language medical dataset by reporting AUC differences between demographic groups.

5.2.4. Bias Detection with Domain-Specific Metrics

Yeo et al. (2025) used the Linguistic Inquiry and Word Count (LIWC-22) tool, which is a 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.

5.2.5. Bias Detection with LLMs as A Judge

Swaminathan et al. (2024) 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.

5.2.6. Medical 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.

Table 2 | Taxonomy of large AI models bias mitigation strategies and description.

Mitigation Taxonomy	Description	Techniques
Pre-Processing	Mitigate bias before model training	Training Data Debugging (Lu et al., 2020; Qian et al., 2022) Projection-based Mitigation (Iskander et al., 2023; Ravfogel et al., 2020)
In-Processing	Mitigate bias during model training	Model Fine-tuning (Devlin et al., 2019; Gallegos et al., 2024) Architecture Modification (Bartl et al., 2020; Han et al., 2021) Loss Function Modification (Huang et al., 2019b; Liu et al., 2019) Decoding Strategy Modification (Gehman et al., 2020; Xu et al., 2020)
Post-Processing	Mitigate bias after model training	Output Ensemble (Chen et al., 2022; Xiao et al., 2024) Output Rewriting (Dhingra et al., 2023; Tokpo and Calders, 2022) Prompt Engineering (Fatemi et al., 2021; Yang et al., 2023)

Pfohl et al. (2024) evaluated Med-PaLM2 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. (2023) 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. (2023) 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.

6. Medical Bias Mitigation

In Section 5, 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 2 presents the descriptions of the three categories of bias mitigation strategies along with examples.

6.1. Pre-Processing Medical Bias Mitigation

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 data (Gallegos et al., 2024). Common techniques include data augmentation (Chakraborty et al., 2021), data filtering (Chakraborty et al., 2020), instance reweighting (Kamiran and Calders, 2012), and synthetic data generation (Chen et al., 2022; Peng et al., 2022; Xiao et al., 2024).

6.2. In-Processing Medical Bias Mitigation

In-processing strategies aim to mitigate bias by modifying model internals during training or fine-tuning (Gallegos et al., 2024). 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.

6.2.1. In-Processing Approaches

Model Fine-tuning. Fine-tuning is a typical practice of in-processing approaches, which involves updating the model parameters via partially re-training the model from extra data with different training settings (Devlin et al., 2019; Gallegos et al., 2024).

Architecture Modification. Architecture-level modifications alter core model components, such as the number and configuration of layers or attention heads (Gallegos et al., 2024; Lauscher et al., 2021). 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 constraints (Gallegos et al., 2024). 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 (Huang et al., 2019b; Liu et al., 2019).

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 (Gehman et al., 2020; Xu et al., 2020).

6.2.2. Practice of In-Processing Approaches

Luo et al. (2024a) 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 is proven to significantly outperform CLIP (Radford et al., 2021) in terms of both performance and fairness.

Jin et al. (2024) 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. (2024) demonstrated that fine-tuning large AI models on local datasets signif-

icantly 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. (2024) 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 gender integration improving fairness by 2.5% and performance by 8.6% in ophthalmology. Zahraei and Shakeri (2024) fine-tuned ChatDoctor using PEFT/LoRA techniques (Hu et al., 2022) 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. (2025a) introduced an adversarial debiasing framework based on variational auto-encoders (VAE), 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 Imran (2025) 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. (2023) 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. (2024) 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. Hastings (Hastings, 2024) 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. (2024) employed fine-tuning on GPT-4o mini with a balanced dataset. While gender 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. (2024b) 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. (2024) 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.

6.3. 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 (Chen et al., 2024b; Gallegos et al., 2024).

6.3.1. Post-Processing Approaches

Output re-writing. 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 response (Chen et al., 2022). This technique leverages the stochasticity of large AI models decoding 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 constraints (White et al., 2023). 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. (2023a) found that asking GPT-3.5 to explain its reasoning prior to providing an answer is able to mitigate gender and racial biases in clinical management of acute coronary syndrome (ACS).

6.3.2. 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 responses (Gallegos et al., 2024).

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. Chain-of-thought prompting (CoT), which encourages models to reason step-by-step, has shown significant promise in reducing biases in clinical decision-making tasks. Poulain et al. (2024b) 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. (2024) 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 Shakeri (2024) 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. (2024) 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. (2024c) 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. (2024) 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. Hastings (2024) 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. (2024) highlighted the variability in prompt engineering effectiveness across models and demographic categories. While some prompts successfully mitigated gender 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 chain-of-thought 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.

7. Available Large AI Models and Datasets

Based on the collected papers, we summarize medical-specific large AI models employed in medical bias research. We also compile publicly available datasets relevant to medical bias detection and mitigation, providing researchers and practitioners with a convenient reference for future work.

7.1. Large AI Models for Medical Bias Research

General-purpose large AI models, such as the GPT family (Achiam et al., 2023; OpenAI, 2025), Claude family (Anthropic, 2025b), and Mistral family (AI, 2025b), are widely adopted in medical bias research. Besides, there have been medical-specific large AI models also been widely adopted for research. To facilitate further research in this direction, we provide basic information and URLs of these medical-specific models in Table 3

Table 3 | Medical-specific large AI models.

Model Name	Models Family	Parameter Size	Year	Open Source	URL
Med-PaLM (Singhal et al., 2023a)	PaLM	>= 175B	2022	No	Med-PaLM
Med-PaLM2 (Singhal et al., 2023b)	PaLM 2	>= 175B	2023	No	Med-PaLM2
PaLMrya-Med (Writer, 2023)	PaLMrya-Med	70B - 175B	2023	Yes	PaLMrya-Med
Meditron (Chen et al., 2023)	LlaMa-2	70B - 175B	2023	Yes	Meditron
OpenBioLLM (Saama AI Labs, 2024)	LlaMa-2	70B - 175B	2023	Yes	OpenBioLLM
PMC-LlaMa (Wu et al., 2024a)	LlaMa	10B - 70B	2023	Yes	PMC-LlaMa
MedAlpaca (Han et al., 2023)	LlaMa	10B - 70B	2023	Yes	MedAlpaca
DoctorGLM (Xiong et al., 2023)	ChatGLM	10B - 70B	2023	Yes	DoctorGLM
Huatuo (Wang et al., 2023e)	LlaMa	10B - 70B	2023	Yes	Huatuo
BioMegatron (Shin et al., 2020)	Megatron-LM	1B - 10B	2021	Yes	BioMegatron
LLaVA-Med (Li and et al., 2023)	LLaVA	1B - 10B	2024	Yes	LLaVA-Med
ChatDoctor (Li et al., 2023c)	LlaMa	1B - 10B	2023	Yes	ChatDoctor
BioMistral (Labrak et al., 2024)	Mistral	1B - 10B	2024	Yes	BioMistral
MedLlaMa-3 (John Snow Labs, 2024)	LlaMa-3	1B - 10B	2024	Yes	MedLlaMa-3
BioBERT (Lee et al., 2020)	BioBERT	< 1B	2019	Yes	BioBERT
ClinicalBERT (Alsentzer et al., 2019)	ClinicalBERT	< 1B	2019	Yes	ClinicalBERT
BioGPT (Luo et al., 2022)	GPT	< 1B	2022	Yes	BioGPT
MedMAE (Gupta et al., 2024)	MAE	< 1B	2023	Yes	MedMAE
MedCLIP (Wang et al., 2022)	CLIP	< 1B	2022	Yes	MedCLIP
PubMedCLIP (Eslami et al., 2023)	CLIP	< 1B	2023	Yes	PubMedCLIP
BiomedCLIP (Zhang et al., 2023b)	CLIP	< 1B	2023	Yes	BiomedCLIP
MentalBERT (Ji et al., 2021)	BERT	< 1B	2021	Yes	MentalBERT
SAMed (Zhang and Liu, 2023)	SAM	< 1B	2023	Yes	SAMed
GatorTron (Yang et al., 2022)	GatotTron	< 1B	2021	Yes	GatorTron
CheXzero (Tiu et al., 2022)	CLIP	< 1B	2022	Yes	CheXzero

These datasets are diverse, covering diverse parameter sizes from <1B to >=175B, covering diverse model families, such as PaLM, LLaMa, ChatGLM, and Mistral, and both open-source and proprietary models, which can facilitate different research purposes.

7.2. Datasets for Medical Bias Research

To support future research, we summarize the datasets employed in current medical bias studies. As shown in Tables 6, 5, and 4, the datasets are organized by modality, including text, image, and multimodal. For each dataset, we provide its source, related diseases, sensitive attributes, and URLs.

Multimodal datasets typically combine image and text data, enabling cross-modal tasks that are essential for evaluating models in integrated scenarios. They support applications such as aligning medical images with corresponding textual descriptions and analyzing interactions between visual and textual information. Image datasets focus on visual data, including X-rays, CT scans, and MRIs, which are critical for disease detection, classification, and localization. Text datasets primarily target natural language processing tasks, such as medical question answering, patient interaction analysis, and mental health assessment.

These datasets are essential resources for training, validating, and benchmarking large AI models in medicine and healthcare applications. They play a critical role in detecting and mitigating biases associated with different demographic factors, including age, race, sex, and socioeconomic status.

Table 4 | Text datasets for medical bias research.

Name	Sources	Related Diseases	Sensitive Attributes	URL
AMQA (Xiao et al., 2025)	United States	General medical dataset	Race, Sex, Socioeconomic Status	AMQA
BiasMD (Zahraei and Shakeri, 2024)	Canada	General medical dataset	Disability, Religion Belief, Sexuality, Socioeconomic Status	BiasMD
BiasMedQA (Schmidgall et al., 2024)	United States	General medical dataset	Cognitive	BiasMedQA
C-SSRS (Gaur et al., 2019)	United States	Mental health	Age, Nationality, Race, Sex	C-SSRS
CAMS (Garg et al., 2022)	Global	Mental health	Age, Nationality, Race, Sex	CAMS
CMExam (Liu et al., 2024a)	China	General medical dataset	Cultural Context, Language	CMExam
CPV (Benkirane et al., 2024)	United States	General medical dataset	Race, Sex	CPV
Cross-Care (Chen et al., 2024a)	United States	General medical dataset	Race, Sex	Cross-Care
DepEmail (Wang et al., 2024d)	United States	Depression, Mental health	Age, Nationality, Race, Sex	DepEmail
DiseaseMatcher (Zahraei and Shakeri, 2024)	Not Specified	General medical dataset	Race, Religion Belief, Socioeconomic Status	DiseaseMatcher
Dreaddit (Turcan and McKeown, 2019)	United States	Mental health	Age, Nationality, Race, Sex	Dreaddit
emrQA (Pampari et al., 2018)	United States	General medical dataset	Age, Nationality, Race, Sex	emrQA
EquityMedQA (Link, 2024)	Not Specified	General medical dataset	Race, Sex, Socioeconomic Status	EquityMedQA
Huatuo-26M (Li et al., 2023b)	China	General medical dataset	Cultural Context, Language	Huatuo-26M
I2B2 (DBMI, Harvard Medical School, 2024a)	United States	General medical dataset	Age, Nationality, Race, Sex	I2B2
IRF (Centers for Medicare & Medicaid Services, 2024)	United States	General medical dataset	Age, Nationality, Race, Sex	IRF
MedQA (Jin et al., 2021)	China, United States	General medical dataset	Cultural Context, Language	MedQA
MedMCQA (Pal et al., 2022)	India	General medical dataset	Cultural Context, Language	MedMCQA
MultiWD (Garg et al., 2024)	Not Specified	Mental health	Age	MultiWD
N2C2 (DBMI, Harvard Medical School, 2024b)	United States	General medical dataset	Age, Race, Sex	N2C2
PMC-Patients (Zhao et al., 2023b)	Not Specified	General medical dataset	Race	PMC-Patients
PubMedQA (Jin et al., 2019)	Not Specified	General medical dataset	Age, Nationality, Race, Sex	PubMedQA
Q-Pain (PhysioNet, 2023)	Not Specified	General medical dataset	Race, Sex	Q-Pain
StressAnnotatedDataset (Mauriello et al., 2021)	United States	General medical dataset	Age, Nationality, Race, Sex	StressAnnotatedDataset
SKMCH&RC (SKMCH&RC, 2024)	Pakistan	General medical dataset	Age, Nationality, Race, Sex	SKMCH&RC
SPIDER (Yu et al., 2018)	United States	General medical dataset	Language	SPIDER
SWMH (AIMH, 2025)	Not Specified	Mental health	Age, Nationality, Race, Sex	SWMH
The Pile (Gao et al., 2021)	United States	General medical dataset	Age, Nationality, Race, Sex	The Pile
TUSC (Vishnubhotla et al., 2022)	Canada, United States	General medical dataset	Age, Nationality, Race, Sex	TUSC

8. Roadmap of LLM Medical Bias Research

8.1. Distributions of Existing Research

Bias in large AI models for medicine and healthcare has attracted growing attention from the academic community, with related research increasing significantly since 2023. In this section, we analyze the distribution of existing studies in this area.

Table 5 | Image datasets for medical bias research.

Name	Sources	Related Diseases	Sensitive Attributes	URL
ADNI-1.5T (Jr. et al., 2008)	United States	Alzheimer's disease	Age, Nationality, Race, Sex, Socioeconomic Status	ADNI-1.5T
CAMELYON17 (Litjens et al., 2018)	Netherlands	Breast cancer	Age, Nationality, Race, Sex	CAMELYON17
ChestDR (Wang et al., 2023a)	China	General Medical Dataset	Age, Nationality, Race, Sex	ChestDR
ChestXray14 (Wang et al., 2017)	United States	Atelectasis, Cardiomegaly, Pleural effusion	Age, Nationality, Race, Sex	ChestXray14
ColonPath (Wang et al., 2023b)	China	Colorectal cancer, Gastrointestinal lesions	Age, Nationality, Race, Sex	ColonPath
HAM10000 (Tschandl et al., 2018)	Australia, Austria	Actinic keratoses, Dermatofibroma, Intraepithelial carcinoma, Vascular lesions	Age, Nationality, Race, Sex	HAM10000
ImageNet (Deng et al., 2009)	Global	General dataset	Age, Cultural Context, Nationality, Race, Sex	ImageNet
MedFMC (Wang et al., 2023c)	China	Colorectal lesions, Diabetic retinopathy, Neonatal jaundice, Pneumonia	Age, Nationality, Race, Sex	MedFMC
Montgomery-County-X-ray (Jaeger et al., 2014)	United States	Tuberculosis (TB)	Age, Nationality, Race, Sex	Montgomery-County-X-ray
NeoJaundice (Wang et al., 2023d)	China	Neonatal jaundice	Age, Nationality, Race, Sex	NeoJaundice
ODIR (Peking University & Grand-Challenge, 2019)	China	Age-related macular degeneration, Cataracts, Diabetes, Glaucoma, Hypertension, Myopia	Age, Nationality, Race, Sex	ODIR
Retino (Rath, 2019)	Not Specified	Diabetic retinopathy	Age, Nationality, Race, Sex	Retino

8.1.1. Medical Scenario

We analyze the distribution of publications related to medical bias across different scenarios. As shown in Figure 4, the 55 collected papers span multiple application contexts: 45 involve clinical decision support, 12 involve medical education, 9 involve medical documentation, and 25 involve patient communication. Because individual studies may address more than one scenario, the total count exceeds 55.

The dominance of clinical decision support indicates that current investigations primarily focus on biases in AI-assisted clinical decision-making, such as diagnostic support and treatment recommendations. This prevalence reflects both the clinical significance and the technological maturity of decision-support systems, which provide clearer evaluation benchmarks and measurable outcomes.

In contrast, studies concerning patient communication, medical documentation, and medical education remain relatively underexplored. A likely reason is that these scenarios involve more subjective, context-dependent, and interaction-oriented tasks, for which standardized datasets and evaluation metrics are less established. As a result, biases in these areas, though potentially more subtle and socially consequential, have received less systematic scrutiny compared to decision-making applications.

8.1.2. Clinical Specialty

Research on medical bias in large AI models spans a wide range of clinical specialties. Figure 5 shows the distribution of the collected papers. Some studies focus on bias in general healthcare scenarios,

Table 6 | Multimodal datasets for medical bias research.

Name	Sources	Related Diseases	Sensitive Attributes	URL
AREDS (National Eye Institute, 2001)	United States	Age-related macular degeneration, Cataracts	Age, Nationality, Race, Sex	AREDS
BRSET (Nakayama et al., 2024)	Brazil	Diabetic retinopathy	Age, Nationality, Race, Sex	BRSET
CANDI (Kennedy et al., 2012)	United States	Neurodevelopmental disorders, Schizophrenia	Age, Nationality, Race, Sex	CANDI
Chest-ImaGenome (Wu et al., 2021)	United States	Atelectasis, Cardiomegaly, Pleural effusion, Pneumonia	Age, Nationality, Race, Sex	Chest-ImaGenome
CheXpert (Irvin et al., 2019)	United States	Atelectasis, Cardiomegaly, Consolidation, Pleural effusion, Pulmonary Edema	Age, Nationality, Race, Sex	CheXpert
COVID-CT-MD (Afshar et al., 2021)	Iran	COVID-19	Age, Nationality, Race, Sex	COVID-CT-MD
FairSeg (Tian et al., 2024)	United States	Glaucoma	Language, Marital Status, Race, Sex	FairSeg
FairVLMed10k (Luo et al., 2024c)	United States	Glaucoma	Language, Marital Status, Race, Sex	FairVLMed10k
GF3300 (Luo et al., 2024b)	United States	Glaucoma	Age, Language, Marital Status, Race, Sex	GF3300
IRCADb (IRCAD, 2010)	France	Liver tumors	Age, Nationality, Race, Sex	IRCADb
KiTS (Heller et al., 2019)	United States	Kidney cancer	Age, Nationality, Race, Sex	KiTS
MIDRC (MIDRC Consortium, 2021)	United States	COVID-19	Age, Nationality, Race, Sex	MIDRC
MIMIC-CXR (Johnson et al., 2019)	United States	Atelectasis, Cardiomegaly, Pleural effusion, Pneumonia	Age, Nationality, Race, Sex	MIMIC-CXR
MIMIC-III (Johnson et al., 2016)	United States	General medical dataset	Age, Nationality, Race, Sex	MIMIC-III
MIMIC-IV (Johnson et al., 2023)	United States	General medical dataset	Age, Nationality, Race, Sex	MIMIC-IV
MSHS (Child and Family Data Archive, 2019)	United States	General medical dataset	Race, Sex, Socioeconomic Staus	MSHS
NEJM Healer Cases (Abdulnour and Kachalia, 2022)	United States	General medical dataset	Age, Nationality, Race, Sex	NEJM Healer Cases
OHTS (Kass et al., 2002)	United States	Glaucoma	Age, Nationality, Race, Sex	OHTS
OL3I (Chaves et al., 2023)	United States	Ischemic heart disease	Age, Nationality, Race, Sex	OL3I
PAD-UFES-20 (Pacheco et al., 2020)	Brazil	Actinic Keratosis, Basal Cell Carcinoma, Melanoma, Nevus, Seborrheic Keratosis, Squamous Cell Carcinoma	Age, Nationality, Race, Sex	PAD-UFES-20
PadChest (Bustos et al., 2020)	Spain	Atelectasis, Cardiomegaly, Emphysema, Fibrosis, Hernia, Nodule, Pleural effusion, Pneumonia, Pneumothorax	Age, Nationality, Race, Sex	PadChest
PAPILA (Kovalyk et al., 2022)	Spain	Glaucoma	Age, Nationality, Race, Sex	PAPILA
VinDr (Nguyen et al., 2022)	Vietnam	Breast cancer, Nodule, Pneumonia, Pneumothorax	Age, Nationality, Race, Sex	VinDr

such as preventive care or general medical question answering (e.g., (Nastasi et al., 2023)), and are grouped as general healthcare.

We find that general healthcare dominates the literature (21 out of 55 papers), likely because widely used benchmarks such as MedQA and AMQA target general medical QA tasks. These datasets provide shared evaluation frameworks, which make bias studies more feasible and comparable across models.

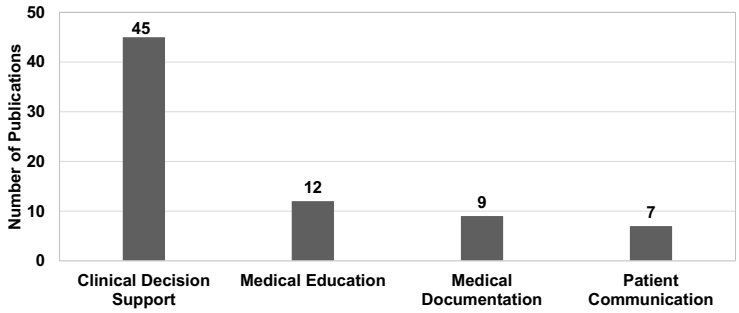

Figure 4 | Distribution of publications by medical scenario.

Figure 5 | Distribution of publications by clinical specialty.

Cardiology (10 papers) and pulmonology (10 papers) also receive notable attention, reflecting both their high clinical impact and the abundance of structured datasets (e.g., ECG, CXR) that enable quantitative bias analysis.

In contrast, endocrinology, neurology, and nephrology remain underrepresented, each with three or fewer studies. This uneven distribution points to a research gap: conditions such as diabetes, chronic kidney disease, and mental health disorders, often affecting marginalized groups, are less examined despite their social and clinical importance. Addressing these gaps is essential for building more inclusive and equitable medical AI systems.

8.1.3. Sensitive Attribute

Figure 6 illustrates the distribution of sensitive attributes considered in medical bias research. We observe that race, sex, and age are the three most frequently studied attributes, appearing in 50, 43, and 17 papers, respectively. This pattern aligns with prior surveys on bias and fairness in general AI domains (Chen et al., 2024b; Hort et al., 2024), where these three attributes are also dominant.

The prevalence of these attributes 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. In contrast, attributes such as socioeconomic status and language appear far less frequently, suggesting that current research may be constrained by the lack of standardized data or clear operational definitions for these factors. This imbalance highlights an important gap — while existing studies predominantly

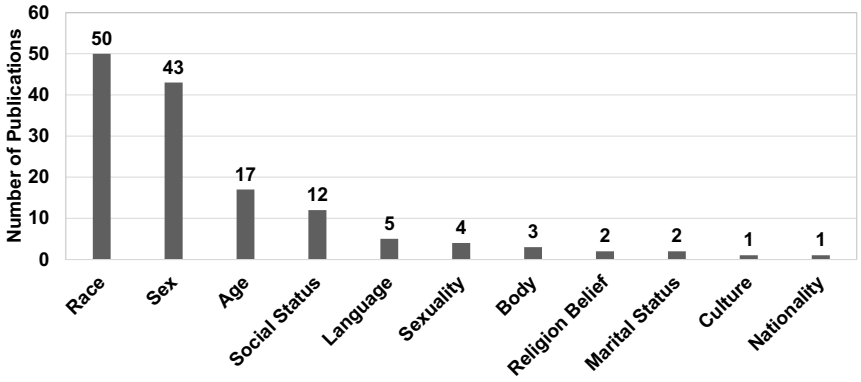

Figure 6 | Distribution of publications by sensitive attribute.

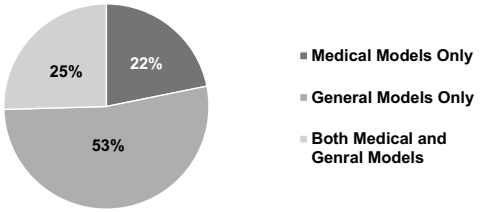

Figure 7 | Model distribution: medical vs general.

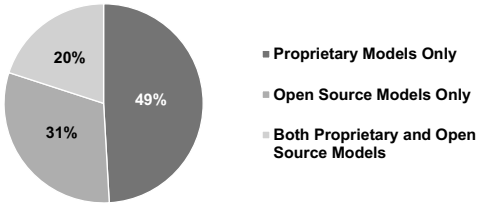

Figure 8 | Model distribution: proprietary vs open source

focus on well-documented demographic biases, contextual and structural factors that can also drive inequities in medical AI systems remain underexplored.

8.1.4. Model Type

The publications included in this survey cover both general-purpose and medical-specific large AI models. Figure 7 presents their distribution. Among all collected papers, 22% focus exclusively on medical foundation models (e.g., Med-PaLM 2, Meditron), while 53% study general-purpose models not originally designed for medical use (e.g., GPT-4, Claude 3.5 Sonnet). The remaining 25% investigate both types of models.

This distribution suggests that, although specialized medical models have recently emerged, general-purpose large AI models still dominate medical bias research. Their prevalence can be attributed to their broader availability, strong baseline performance, and frequent use as reference systems across both general and medical domains.

We further analyze the models in terms of source accessibility. As shown in Figure 8, 69% of the papers (i.e., 49% + 20%) conduct experiments on proprietary models, such as the GPT and Claude series. Notably, 64% of these studies focus on investigating the medical bias of GPT models (e.g., GPT-3.5 and GPT-4). In contrast, 51% of the collected works (i.e., 31% + 20%) examine open-source models, including the LLaMA and Mistral families.

The overall trend reflects the dual influence of closed- and open-source ecosystems. Proprietary models such as GPT continue to shape the research frontier due to their strong capabilities and easy access via APIs, whereas open-source models, though offering greater flexibility and transparency, remain less accessible because of high computational demands. This imbalance highlights a practical

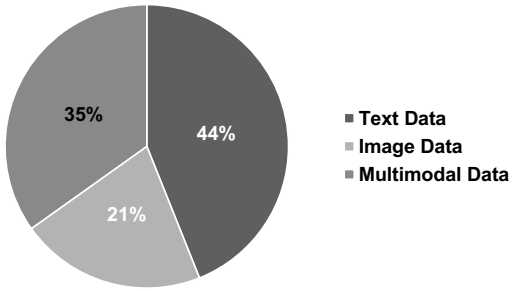

Figure 9 | Distribution of publications by data type.

barrier for bias auditing and underscores the need for more open, reproducible resources in medical AI fairness research.

8.1.5. Data Type

As shown in Figure 9, the distribution of datasets in the collected papers exhibits a clear imbalance across data modalities. Text data dominates (44%), reflecting the widespread adoption of large AI models for text-related tasks such as clinical documentation, electronic health record analysis, and medical question answering. This prevalence suggests that current studies are largely driven by the maturity of natural language processing pipelines and the accessibility of textual benchmarks.

Image datasets account for 21%, despite the central role of medical imaging in clinical decision-making. Their relatively limited use may stem from privacy concerns, annotation costs, and the requirement for domain expertise in image interpretation.

Meanwhile, multimodal datasets represent 35%, indicating a growing effort to integrate heterogeneous information sources such as text, images, and structured signals (e.g., laboratory results, physiological waveforms). This shift toward multimodality is exemplified by recent models like LLaVA-Med, which underscore a broader trend toward holistic patient modeling that better reflects real-world clinical contexts.

Overall, the distribution highlights a methodological concentration on text-based research, while multimodal integration remains an emerging but promising direction for achieving more comprehensive and equitable medical AI systems.

8.1.6. Venue

Figure 10 shows the venue distribution of existing research on medical bias in large AI models. As this line of research lies at the intersection of artificial intelligence (AI) and medicine, both medical and AI/Computer Science (CS) venues serve as key publication outlets. Our analysis reveals that 38% of the studies are published in medical journals and conferences (e.g., Nature Medicine, Lancet Digital Health), while 13

Notably, a substantial proportion of studies, 33% on arXiv and 13% on MedRxiv, are disseminated through preprint platforms. This trend reflects both the rapidly evolving nature of large AI models and the slow publication cycles of traditional venues, prompting researchers to share their findings more promptly via open repositories.

The observed distribution suggests a growing convergence between AI and medical communities, yet also indicates that much of the discourse on bias in medical AI is still emerging.

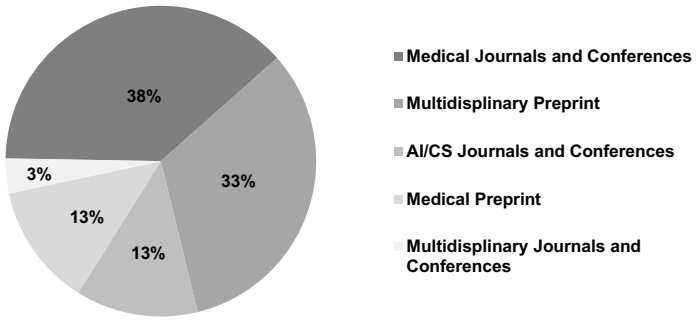

Figure 10 | Distribution of publications by venue.

8.2. Open Problems and Research Opportunities

Although there has been increasing attention to bias in large AI models for medicine and healthcare, the field remains fragmented and faces many unresolved challenges. Below we identify several key open problems and outline promising research opportunities to guide future work.

8.2.1. Lack of Unified Foundations for Medical Fairness

Current studies 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 (Bommasani et al., 2021; Guo et al., 2024; Pfohl et al., 2024):

- **Unjustified Bias**: Disparities arising from historical discrimination, unequal access, or flawed data collection that are not rooted in a biologically or clinically meaningful difference.
- **Clinically Relevant Differences**: Variations in disease presentation, risk, or treatment response that are grounded in evidence and are essential for providing equitable, personalized care.

In this survey, we synthesize existing research on medical bias and define medical bias as *any systematic error or prejudice in the output of an AI system that disadvantages certain individuals or groups*. This definition directly addresses the core problem by shifting the focus from purely statistical disparities to the nature and consequence of the disparity.

Research opportunity: We call for more context-aware medical fairness frameworks that formally encode clinical context, causal assumptions, and ethical criteria. This includes distinguishing algorithmic bias from appropriate personalized medicine and defining fairness relative to clinical guidelines or equity-informed standards of care. The purpose is to have a unified foundation to make the distinguish between unjustified bias and clinically relevant differences easier.

8.2.2. Insufficient Datasets and Evaluation Benchmarks

High-quality datasets and benchmarks are essential for improving and understanding the performance and trustworthiness of large AI model–based systems in medicine and healthcare.

Section 7 introduces medical text datasets, medical image datasets, and medical multimodal datasets referenced in the studies in our paper collection. Nevertheless, they are often small-scale, require human judgment, and lack coverage of intersectional attributes, diverse health systems, and

multimodal workflows. Xiao et al. (2025) proposed the AMQA benchmark, which enables automatic bias detection, but it only considers a limited set of sensitive attributes including race, sex, and socioeconomic status, using counterfactual pairs such as White vs. Black, Male vs. Female, and High-Income vs. Low-Income. Many other demographic subgroups remain unexplored.

Research opportunity: We call for more work on constructing scalable, standardized benchmarks for medical bias evaluation, pre-training, as well as fine-tuning based bias mitigation across modalities, languages, and demographic contexts. There is also a need of platform building for model developers and providers to work with human experts in the user community on identifying and documenting any errors or biased outputs which the deployed AI models might produce (Harrer, 2023).

8.2.3. Lack of Methods on Rigorous Automatic Bias Detection

A significant gap in the current ecosystem is the lack of methods for bias detection on the fly. Most existing fairness techniques focus on static benchmarking, which is useful for retrospectively understanding and ranking an AI model's overall bias on a fixed dataset. However, this approach is insufficient for proactively exposing the myriad ways bias can manifest in dynamic, real-world clinical settings. To truly safeguard against harm, we need a paradigm shift towards software testing methodologies that can aggressively surface bias issues before deployment.

This requires the development of sophisticated test generation methods, such as fuzzing (Manès et al., 2019), metamorphic testing (Chen et al., 2018; Segura et al., 2016), and differential testing (McKeeman, 1998) offer valuable opportunities. Fuzzing can generate a wide variety of inputs, including those that mimic diverse clinical scenarios, to reveal unexpected behaviors and potential bias in model outputs (Manès et al., 2019). Metamorphic testing, by leveraging known relationships between input transformations and expected output invariance, can serve as a powerful tool to assess whether similar clinical queries yield consistent and fair responses (Chen et al., 2018; Segura et al., 2016). Differential testing, which involves comparing outputs across different model versions or configurations when provided with controlled variations in input, may further illuminate discrepancies attributable to bias (McKeeman, 1998).

Furthermore, these techniques must be automatic and scalable to rigorously evaluate complex models across thousands of simulated scenarios, ensuring that bias is not merely measured, but actively hunted and eliminated.

8.2.4. Missing Real-World and Continuous Validation

Most existing studies analyse 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-makers (Singh et al., 2023).

Research opportunity: More work is needed for continuous, longitudinal monitoring systems, such as to develop statistical methods and infrastructure to track fairness metrics over time alongside model performance, and to build realistic simulation platforms that model the interaction loops between AI tools, clinicians, and patient populations.

8.2.5. Inadequate Representation and Global Health Inequity

Current research overwhelmingly focuses on bias in U.S. and European populations, with limited attention to global health disparities. Datasets rarely include underrepresented groups such as children, low-resource regions, rare disease patients, or multilingual settings. This issue is not merely a data gap, it actively propagates and amplifies existing global health inequities. Models trained on homogenous, western data perform poorly when deployed in different contexts, leading to misdiagnosis, inadequate treatment recommendations, and a reinforcement of the disparities.

Research opportunity: A critical and urgent need exists to establish comprehensive, global fairness datasets, with a prioritized focus on data collected from low- and middle-income countries. We call for more work on assessing whether AI models interpret symptoms and patient histories expressed in diverse languages and cultural frameworks, and investigate model bias in multilingual clinical scenarios.

8.2.6. Lack of Studies on the Trade-off between Fairness and Accuracy

It is widely acknowledged that fairness and accuracy in ML models often exhibit a trade-off (Chen et al., 2024c, 2025a,b). 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 the mitigation of bias interacts with other crucial alignment goals, such as safety (avoiding harmful outputs) and explainability (providing interpretable reasoning).

Research opportunity: We call for more studies on understanding the trade-offs between fairness and other requirements regarding large AI models for health. There is also an opportunity to developing multi-objective optimization methods.

9. Conclusion

This survey provides a comprehensive investigation of research on detecting and mitigating bias in large AI models for healthcare. By synthesizing existing literature, it presents a clear and concise definition of bias specific to large AI models in healthcare and systematically introduces a framework for bias evaluation. Furthermore, the survey proposes a structured taxonomy categorizing existing approaches, offering a clear understanding of the current state of research on bias detection and mitigation. By summarizing existing studies and analyzing research trends and distributions, this survey identifies critical open problems and highlights promising directions for future research. Additionally, by compiling a detailed index of relevant datasets and large AI models, it serves as a valuable resource for researchers and practitioners entering this emerging field.

References

R. E. E. Abdulnour and A. Kachalia. Deliberate practice at the virtual bedside to improve diagnostic reasoning. *NEJM*, 2022. URL https://healer.lecturio.com/app/. Describes NEJM Healer approach.

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

P. Afshar, S. Heidarian, N. Enshaei, et al. Covid-ct-md: Covid-19 computed tomography scan dataset applicable in machine learning and deep learning. *Scientific Data*, 2021. doi: 10.1038/s41597-021-00900-3. URL https://github.com/ShahinSHH/COVID-CT-MD.

A. Agrawal. Fairness in ai-driven oncology: Investigating racial and gender biases in large language models. *Cureus*, 16(9), 2024.

M. AI. The llama 4 herd: The beginning of a new era of natively multimodal intelligence. Meta AI Blog, Apr. 2025a. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

M. AI. La Plateforme - frontier LLMs | Mistral AI — mistral.ai. https://mistral.ai/products/la-plateforme, 2025b. [Accessed 04-10-2025].

AIMH. Swmh dataset card. Hugging Face Dataset, 2025. URL https://huggingface.co/datasets/AIMH/SWMH.

N. Akhras, F. Antaki, F. Mottet, O. Nguyen, S. Sawhney, S. Bajwah, and J. M. Davies. Large language models perpetuate bias in palliative care: development and analysis of the palliative care adversarial dataset (pcad). *arXiv preprint arXiv:2502.08073*, 2025.

J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelovic, M. Botvinick, M. Dehghani, A. Clark, and et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS) 35*, pages 23716–23729, 2022.

E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical NLP Workshop*, 2019.

Anthropic. Claude opus 4 & claude sonnet 4 — system card. Technical report, Anthropic, July 2025a. URL https://www.anthropic.com/news/claude-4.

Anthropic. Home - Claude Docs — docs.claude.com. https://docs.claude.com/en/home, 2025b. [Accessed 04-10-2025].

D. U. Apakama, K.-A.-N. Nguyen, D. Hyppolite, S. Soffer, A. Mudrik, E. Ling, A. Moses, I. Temnycky, A. Glasser, R. Anderson, et al. Identifying and characterizing bias at scale in clinical notes using large language models. *medRxiv*, pages 2024–10, 2024.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

M. Bartl, M. Nissim, and A. Gatt. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. *arXiv preprint arXiv:2010.14534*, 2020.

K. Benkirane, J. Kay, and M. Perez-Ortiz. How can we diagnose and treat bias in large language models for clinical decision-making? *arXiv preprint arXiv:2410.16574*, 2024.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 2020. doi: 10.1016/j.media.2020.101797.

Centers for Medicare & Medicaid Services. Inpatient rehabilitation facility (irf) compare / provider data catalog. Website, 2024. URL https://data.cms.gov/provider-data/dataset/v9e4-nwhh.

J. Chakraborty, S. Majumder, Z. Yu, and T. Menzies. Fairway: A way to build fair ml software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 654–665, 2020.

J. Chakraborty, S. Majumder, and T. Menzies. Bias in machine learning software: why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440, 2021.

K. Chansiri, X. Wei, and K. H. B. Chor. Addressing gender bias: A fundamental approach to ai in mental health. In *2024 5th International Conference on Big Data Analytics and Practices (IBDAP)*, pages 107–112. IEEE, 2024.

J. M. Z. Chaves, B. Patel, A. Chaudhari, et al. Opportunistic assessment of ischemic heart disease risk using abdominopelvic ct and medical record data: A multimodal explainable ai approach. *NPJ Digital Medicine*, 2023. URL https://stanfordaimi.azurewebsites.net/datasets/3263e34a-252e-460f-8f63-d585a9bfecfc. Releases the OL3I dataset.

S. Chen, J. Wu, S. Hao, D. Khashabi, D. Roth, et al. Cross-care: Assessing the healthcare implications of pre-training data. In *NeurIPS 2024 Datasets and Benchmarks*, 2024a. URL https://openreview.net/forum?id=9y7ebdAEiv.

T. Y. Chen, F.-C. Kuo, H. Liu, P.-L. Poon, D. Towey, T. Tse, and Z. Q. Zhou. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys (CSUR)*, 51(1):1–27, 2018.

Z. Chen, J. Zhang, F. Sarro, and M. Harman. MAAT: A novel ensemble approach to addressing fairness and performance bugs for machine learning software. In *The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, 2022.

Z. Chen, A. Tang, T. Vu, and et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

Z. Chen, J. M. Zhang, M. Hort, M. Harman, and F. Sarro. Fairness testing: A comprehensive survey and analysis of trends. *ACM Transactions on Software Engineering and Methodology*, 33(5):1–59, 2024b.

Z. Chen, J. M. Zhang, F. Sarro, and M. Harman. Fairness improvement with multiple protected attributes: How far are we? In *Proceedings of the IEEE/ACM 46th international conference on software engineering*, pages 1–13, 2024c.

Z. Chen, X. Li, J. M. Zhang, F. Sarro, and Y. Liu. Diversity drives fairness: Ensemble of higher order mutants for intersectional fairness of machine learning software. In *47th IEEE/ACM International Conference on Software Engineering, ICSE 2025, Ottawa, ON, Canada, April 26 - May 6, 2025*, pages 743–755, 2025a.

Z. Chen, X. Li, J. M. Zhang, W. Sun, Y. Xiao, T. Li, Y. Lou, and Y. Liu. Software fairness dilemma: Is bias mitigation a zero-sum game? *Proc. ACM Softw. Eng.*, 2(FSE):1780–1801, 2025b.

Child and Family Data Archive. Migrant and seasonal head start study (mshs), united states, 2017-2018. https://www.childandfamilydataarchive.org/cfda/archives/cfda/studies/37348, 2019.

J. Czum and S. Parr. Bias in foundation models: primum non nocere or caveat emptor? *Radiology: Artificial Intelligence*, 5(6):e230384, 2023.

DBMI, Harvard Medical School. i2b2 data portal. Website, 2024a. URL https://portal.dbmi.hms.harvard.edu/.

DBMI, Harvard Medical School. n2c2 nlp research datasets. Website, 2024b. URL https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/.

DeepSeek. Deepseek-v3.1. API doc, Sept. 2025. URL https://api-docs.deepseek.com/news/news250821.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

H. Dhingra, P. Jayashanker, S. Moghe, and E. Strubell. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *arXiv preprint arXiv:2307.00101*, 2023.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Y. Du, J.-t. Huang, J. Zhao, and L. Lin. Faircoder: Evaluating social bias of llms in code generation. *arXiv preprint arXiv:2501.05396*, 2025.

T. A. D'Antonoli, A. Stanzione, C. Bluethgen, F. Vernuccio, L. Ugga, M. E. Klontzas, R. Cuocolo, R. Cannella, and B. Koçak. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagnostic and Interventional Radiology*, 30(2):80, 2024.

S. Eslami, F. Benitez-Quiroz, and A. M. Martinez. How much does clip benefit visual question answering in the medical domain? In *Findings of EACL*, 2023. URL https://aclanthology.org/2023.findings-eacl.88.pdf.

Z. Fatemi, C. Xing, W. Liu, and C. Xiong. Improving gender fairness of pre-trained language models without catastrophic forgetting. *arXiv preprint arXiv:2110.05367*, 2021.

H. Fayyaz, R. Poulain, and R. Beheshti. Enabling scalable evaluation of bias patterns in medical llms. *arXiv preprint arXiv:2410.14763*, 2024.

I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024.

L. Gao, S. Biderman, S. Black, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2021. URL https://arxiv.org/abs/2101.00027.

Y. Gao, R. Li, E. Croxford, S. Tesch, D. To, J. Caskey, B. W. Patterson, M. M. Churpek, T. Miller, D. Dligach, et al. Large language models and medical knowledge grounding for diagnosis prediction. *medRxiv*, pages 2023–11, 2023.

M. Garg, M. Kokkodis, A. Khan, L.-P. Morency, M. D. Choudhury, M. S. A. Hussain, et al. Cams: Causes for mental health problems—a new task and dataset. In *Proceedings of LREC*, 2022.

M. Garg et al. Multiwd: Multi-label wellness dimensions in social media posts. GitHub repository and preprint, 2024. URL https://github.com/drmuskangarg/MultiWD.

M. Gaur, A. Alambo, V. Shalin, U. Kursuncu, K. Thirunarayan, A. Sheth, J. Pathak, et al. Reddit c-ssrs suicide dataset. Zenodo, 2019. URL https://zenodo.org/records/5920080.

S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

B. Glocker, C. Jones, M. Roschewitz, and S. Winzeck. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, 5(6):e230060, 2023.

E. Goh, B. Bunning, E. Khoong, R. Gallo, A. Milstein, D. Centola, and J. H. Chen. Chatgpt influence on medical decision-making, bias, and equity: a randomized study of clinicians evaluating clinical vignettes. *Medrxiv*, 2023.

E. Goh, B. Bunning, E. C. Khoong, R. J. Gallo, A. Milstein, D. Centola, and J. H. Chen. Physician clinical decision modification and bias assessment in a randomized controlled trial of ai assistance. *Communications Medicine*, 5(1):59, 2025.

Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, H. Li, M. Qiu, and S. S. Liu. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*, 2024.

A. Gupta, I. I. Osman, M. S. Shehata, and J. W. Braun. Medmae: A self-supervised backbone for medical imaging tasks. *arXiv preprint arXiv:2407.14784*, 2024.

T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Figueroa, A. Löser, D. Truhn, and K. K. Bressem. Medalpaca – an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

X. Han, T. Baldwin, and T. Cohn. Balancing out bias: Achieving fairness through balanced training. *arXiv preprint arXiv:2109.08253*, 2021.

J. J. Hanna, A. D. Wakene, C. U. Lehmann, and R. J. Medford. Assessing racial and ethnic bias in text generation for healthcare-related tasks by chatgpt. *MedRxiv*, 2023.

J. J. Hanna, A. D. Wakene, A. O. Johnson, C. U. Lehmann, and R. J. Medford. Assessing racial and ethnic bias in text generation by large language models for health care–related tasks: Cross-sectional study. *Journal of Medical Internet Research*, 27:e57257, 2025.

S. Harrer. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90, 2023.

S. A. Hasheminasab, F. Jamil, M. U. Afzal, A. H. Khan, S. Ilyas, A. Noor, S. Abbas, H. N. Cheema, M. U. Shabbir, I. Hameed, 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. *medRxiv*, pages 2024–06, 2024.

J. Hastings. Preventing harm from non-conscious bias in medical generative ai. *The Lancet Digital Health*, 6(1):e2–e3, 2024.

Y. He, F. Huang, X. Jiang, Y. Nie, M. Wang, J. Wang, and H. Chen. Foundation model for advancing healthcare: Challenges, opportunities, and future directions. *arXiv preprint arXiv:2404.03264*, 2024.

N. Heller, N. Sathianathen, A. Kalapara, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019. URL https://kits-challenge.org/.

M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2):1–52, 2024.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

J.-t. Huang, J. Qin, J. Zhang, Y. Yuan, W. Wang, and J. Zhao. Visbias: Measuring explicit and implicit social biases in vision language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17981–18004, 2025a.

J.-t. Huang, Y. Yan, L. Liu, Y. Wan, W. Wang, K.-W. Chang, and M. R. Lyu. Where fact ends and fairness begins: Redefining ai bias evaluation through cognitive biases. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10974–10993, 2025b.

K. Huang, J. Altosaar, and R. Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019a.

P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, and P. Kohli. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*, 2019b.

X. Huang, H. Zhang, X. Zhou, M. A. Babar, and S. Yang. Synthesizing qualitative research in software engineering: A critical review. In *Proceedings of the 40th international conference on software engineering*, pages 1207–1218, 2018.

IRCAD. 3d-ircadb-01: Liver segmentation dataset. https://www.ircad.fr/research/data-sets/liver-segmentation-3d-ircadb-01/, 2010.

J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

S. Iskander, K. Radinsky, and Y. Belinkov. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. *arXiv preprint arXiv:2305.10204*, 2023.

N. Ito, S. Kadomatsu, M. Fujisawa, K. Fukaguchi, R. Ishizawa, N. Kanda, D. Kasugai, M. Nakajima, T. Goto, and Y. Tsugawa. The accuracy and potential racial and ethnic biases of gpt-4 in the diagnosis and triage of health conditions: evaluation study. *JMIR Medical Education*, 9:e47532, 2023.

S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 2014. URL https://qims.amegroups.org/article/view/5132.

S. Jalali and C. Wohlin. Systematic literature studies: database searches vs. backward snowballing. In *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*, pages 29–38, 2012.

S. Ji, E. Camacho-Collados, N. Aletras, Y. Yang, and et al. Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*, 2021. URL https://arxiv.org/abs/2110.15621.

Y. Ji, W. Ma, S. Sivarajkumar, H. Zhang, E. M. Sadhu, Z. Li, X. Wu, S. Visweswaran, and Y. Wang. Mitigating the risk of health inequity exacerbated by large language models. *npj Digital Medicine*, 8(1):246, 2025.

D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP-IJCNLP*, 2019.

R. Jin, Z. Xu, Y. Zhong, Q. Yao, Q. Dou, S. K. Zhou, and X. Li. Fairmedfm: fairness benchmarking for medical imaging foundation models. *arXiv preprint arXiv:2407.00983*, 2024.

John Snow Labs. Jsl-medllama-3-8b-v2.0. https://huggingface.co/johnsnowlabs/JSL-MedLlama-3-8B-v2.0, 2024.

A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

A. E. W. Johnson, L. Bulgarelli, L. Shen, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 2023. doi: 10.1038/s41597-022-01899-x. URL https://physionet.org/content/mimiciv/3.1/.

C. Jones, D. C. Castro, F. De Sousa Ribeiro, O. Oktay, M. McCradden, and B. Glocker. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6 (2):138–146, 2024.

C. R. J. Jr., M. W. Weiner, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 2008.

K. S. Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048, 2024.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

P. K. Kanithi, C. Christophe, M. A. Pimentel, T. Raha, N. Saadi, H. Javed, S. Maslenkova, N. Hayat, R. Rajan, and S. Khan. Medic: Towards a comprehensive framework for evaluating llms in clinical applications. *arXiv preprint arXiv:2409.07314*, 2024.

M. A. Kass, D. K. Heuer, E. J. Higginbotham, et al. The ocular hypertension treatment study: A randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Archives of Ophthalmology*, 2002. doi: 10.1001/archopht.120.6.701.

Y. H. Ke, R. Yang, S. A. Lie, T. X. Y. Lim, H. R. Abdullah, D. S. W. Ting, and N. Liu. Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias. *arXiv preprint arXiv:2401.14589*, 2024.

D. N. Kennedy, C. Haselgrove, S. M. Hodge, et al. Candishare: A resource for pediatric neuroimaging data. *Neuroinformatics*, 2012. URL https://www.nitrc.org/projects/candi_share/.

M. O. Khan, M. M. Afzal, S. Mirza, and Y. Fang. How fair are medical imaging foundation models? In *Machine Learning for Health (ML4H)*, pages 217–231. PMLR, 2023.

J. Kim, Z. R. Cai, M. L. Chen, J. F. Simard, and E. Linos. Assessing biases in medical decisions via clinician and ai chatbot responses to patient vignettes. *JAMA Network Open*, 6(10):e2338050–e2338050, 2023.

J. Kim, K. G. Leonte, M. L. Chen, J. B. Torous, E. Linos, A. Pinto, and C. I. Rodriguez. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *NPJ Digital Medicine*, 7(1):193, 2024.

Y. Kim, H. Jeong, S. Chen, S. S. Li, M. Lu, K. Alhamoud, J. Mun, C. Grau, M. Jung, R. R. Gameiro, et al. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*, pages 2025–02, 2025.

Kimi, Y. Bai, Y. Bao, G. Chen, J. Chen, N. Chen, R. Chen, Y. Chen, Y. Chen, Y. Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

O. Kovalyk, B. Remeseiro, M. Ortega, et al. Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data*, 2022. doi: 10.1038/s41597-022-01388-1. URL https://figshare.com/articles/dataset/PAPILA/14798004/2.

I. Ktena, O. Wiles, I. Albuquerque, S.-A. Rebuffi, R. Tanno, A. G. Roy, S. Azizi, D. Belgrave, P. Kohli, T. Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, pages 1–8, 2024.

Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

A. Lauscher, T. Lueken, and G. Glavaš. Sustainable modular debiasing of language models. *arXiv preprint arXiv:2109.03646*, 2021.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

C. Li and et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

H. Li, J. T. Moon, S. Purkayastha, L. A. Celi, H. Trivedi, and J. W. Gichoya. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5(6):e333–e335, 2023a.

J. Li, X. Wang, X. Wu, Z. Zhang, X. Xu, J. Fu, P. Tiwari, X. Wan, and B. Wang. Huatuo-26m, a large-scale chinese medical qa dataset. arXiv:2305.01526 and dataset card, 2023b.

M. Li, H. Chen, Y. Wang, T. Zhu, W. Zhang, K. Zhu, K.-F. Wong, and J. Wang. Understanding and mitigating the bias inheritance in llm-based data augmentation on downstream tasks. *arXiv preprint arXiv:2502.04419*, 2025.

Y. Li, Z. Li, K. Zhang, R. Dan, and Y. Zhang. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023c. URL https://arxiv.org/abs/2303.14070.

B. Lin, N. Cassee, A. Serebrenik, G. Bavota, N. Novielli, and M. Lanza. Opinion mining for software development: a systematic literature review. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(3):1–41, 2022.

K. Link. Equitymedqa dataset card. Hugging Face Dataset, 2024. URL https://huggingface.co/datasets/katielink/EquityMedQA.

G. Litjens, P. Bandi, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer metastases: the camelyon dataset. *GigaScience*, 2018. URL https://academic.oup.com/gigascience/article/7/6/giy065/5026175. Article ID giy065.

H. Liu, J. Dacon, W. Fan, H. Liu, Z. Liu, and J. Tang. Does gender matter? towards fairness in dialogue systems. *arXiv preprint arXiv:1910.10486*, 2019.

J. Liu, P. Zhou, Y. Hua, D. Chong, Z. Tian, A. Liu, H. Wang, C. You, Z. Guo, L. Zhu, et al. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36, 2024a.

L. Liu, X. Yang, J. Lei, X. Liu, Y. Shen, Z. Zhang, P. Wei, J. Gu, Z. Chu, Z. Qin, et al. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*, 2024b.

Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, and H. Xu. Entity recognition from clinical texts via recurrent neural network. *BMC medical informatics and decision making*, 17(Suppl 2):67, 2017.

K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pages 189–202, 2020.

R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

Y. Luo, M. Shi, M. O. Khan, M. M. Afzal, H. Huang, S. Yuan, Y. Tian, L. Song, A. Kouhana, T. Elze, et al. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301, 2024a.

Y. Luo, Y. Tian, F. Liu, et al. Harvard glaucoma fairness: A retinal nerve disease dataset for fairness learning and fair identity normalization. *IEEE Transactions on Artificial Intelligence*, 2024b. URL https://ophai.hms.harvard.edu/datasets/harvard-gf3300/. Dataset page: Harvard GF3300.

Y. Luo, Y. Tian, J. Zhang, et al. Fairclip: Harnessing fairness in vision-language learning. In *CVPR*, 2024c. URL https://ophai.hms.harvard.edu/datasets/harvard-fairvlmed10k. Introduces Harvard-FairVLMed10k dataset.

V. J. Manès, H. Han, C. Han, S. K. Cha, M. Egele, E. J. Schwartz, and M. Woo. The art, science, and engineering of fuzzing: A survey. *IEEE Transactions on Software Engineering*, 47(11):2312–2331, 2019.

M. L. Mauriello, E. T. Lincoln, G. Hon, D. Simon, D. Jurafsky, and P. E. Paredes. Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In *CHI '21 Extended Abstracts*, 2021. doi: 10.1145/3411763.3451799. URL https://github.com/PervasiveWellbeingTech/Stress-Annotated-Dataset-SAD.

D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni, et al. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164*, 2023.

W. M. McKeeman. Differential testing for software. *Digital Technical Journal*, 10(1):100–107, 1998.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

MIDRC Consortium. Medical imaging and data resource center (midrc). https://www.midrc.org/midrc-data, 2021.

MistralAI. Mistral medium 3.1. Meta AI Blog, Aug. 2025. URL https://docs.mistral.ai/models/mistral-medium-3-1-25-08.

N. Munia and A. A. Z. Imran. Prompting medical vision-language models to mitigate diagnosis bias by generating realistic dermoscopic images. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2025.

L. F. Nakayama, F. de Souza, G. Vieira, et al. Brset: A brazilian multilabel ophthalmological dataset of fundus photographs. *PLOS Digital Health*, 2024. doi: 10.1371/journal.pdig.0000454. URL https://physionet.org/content/brazilian-ophthalmological/1.0.0/.

A. J. Nastasi, K. R. Courtright, S. D. Halpern, and G. E. Weissman. A vignette-based evaluation of chatgpt's ability to provide appropriate and equitable medical advice across care contexts. *Scientific Reports*, 13(1):17885, 2023.

National Eye Institute. Nei age-related eye disease study (areds). https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1, 2001.

Z. A. Nazi and W. Peng. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI, 2024.

R. O. Ness, K. Matton, H. Helm, S. Zhang, J. Bajwa, C. E. Priebe, and E. Horvitz. Medfuzz: Exploring the robustness of large language models in medical question answering. *arXiv preprint arXiv:2406.06573*, 2024.

H. Q. Nguyen, K. Lam, L. T. Le, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 2022. doi: 10.1038/s41597-022-01498-w. URL https://vindr.ai/datasets.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

M. Omar, S. Soffer, R. Agbareia, N. L. Bragazzi, D. U. Apakama, C. R. Horowitz, A. W. Charney, R. Freeman, B. Kummer, B. S. Glicksberg, et al. Sociodemographic biases in medical decision making by large language models. *Nature Medicine*, pages 1–9, 2025.

J. A. Omiye, J. C. Lester, S. Spichak, V. Rotemberg, and R. Daneshjou. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195, 2023.

J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou. Large language models in medicine: the potentials and pitfalls: a narrative review. *Annals of Internal Medicine*, 177(2):210–220, 2024.

OpenAI. Introducing chatgpt, 2022. URL https://openai.com/blog/chatgpt/. Accessed: 2024-07-15.

OpenAI. Introducing GPT-5 — openai.com. https://openai.com/index/introducing-gpt-5/, 2025. [Accessed 03-10-2025].

A. G. C. Pacheco, G. R. Lima, A. S. Salomão, et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 2020. doi: 10.1016/j.dib.2020.106221. URL https://data.mendeley.com/datasets/zr7vgbcyr2/1.

S. Pahune and N. Rewatkar. Healthcare: A growing role for large language models and generative ai. *International Journal for Research in Applied Science and Engineering Technology*, 11(8):2288–2301, 2023.

A. Pal, L. K. Umapathi, and M. Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

A. Pampari, P. Raghavan, J. Liang, and J. Peng. emrqa: A large corpus for question answering on electronic medical records. In *EMNLP*, 2018. URL https://aclanthology.org/D18-1187/.

Y.-J. Park, A. Pillai, J. Deng, E. Guo, M. Gupta, M. Paget, and C. Naugler. Assessing the research landscape and clinical utility of large language models: A scoping review. *BMC Medical Informatics and Decision Making*, 24(1):72, 2024.

Peking University & Grand-Challenge. Ocular disease intelligent recognition (odir-2019) dataset. https://odir2019.grand-challenge.org/dataset/, 2019.

K. Peng, J. Chakraborty, and T. Menzies. Fairmask: Better fairness via model-based rebalancing of protected attributes. *IEEE Transactions on Software Engineering*, 2022.

S. R. Pfohl, H. Cole-Lewis, R. Sayres, D. Neal, M. Asiedu, A. Dieng, N. Tomasev, Q. M. Rashid, S. Azizi, N. Rostamzadeh, et al. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, pages 1–11, 2024.

PhysioNet. Q-pain: Evaluation datasets for clinical decision support systems (physionet). Dataset on PhysioNet, 2023. URL https://physionet.org/content/q-pain/1.0.0/.

R. Poulain, H. Fayyaz, and R. Beheshti. Aligning (medical) llms for (counterfactual) fairness. *arXiv preprint arXiv:2408.12055*, 2024a.

R. Poulain, H. Fayyaz, and R. Beheshti. Bias patterns in the application of llms for clinical decision support: A comprehensive study. *arXiv preprint arXiv:2404.15149*, 2024b.

R. Qian, C. Ross, J. Fernandes, E. Smith, D. Kiela, and A. Williams. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*, 2022.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training, 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

M. Rani, B. K. Mishra, D. Thakker, M. Babar, W. Jones, and A. Din. Biases and trustworthiness challenges with mitigation strategies for large language models in healthcare. In *2024 International Conference on IT and Industrial Technologies (ICIT)*, pages 1–6. IEEE, 2024.

S. R. Rath. Diabetic retinopathy 224x224 (2019 data). https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-2019-data, 2019.

S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*, 2020.

Saama AI Labs. Openbiollm-70b (llama3): An open-source biomedical language model. https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B, 2024.

A. W. Salehi, S. Khan, G. Gupta, B. I. Alabduallah, A. Almjally, H. Alsolai, T. Siddiqui, and A. Mellit. A study of cnn and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability*, 15(7):5930, 2023.

S. Schmidgall, C. Harris, I. Essien, D. Olshvang, T. Rahman, J. W. Kim, R. Ziaei, J. Eshraghian, P. Abadir, and R. Chellappa. Addressing cognitive bias in medical language models. *arXiv preprint arXiv:2402.08113*, 2024.

R. Schnepper, N. Roemmel, R. Schaefert, L. Lambrecht-Walzinger, G. Meinlschmidt, 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 Mental Health*, 12(1):e57986, 2025.

S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Cortés. A survey on metamorphic testing. *IEEE Transactions on software engineering*, 42(9):805–824, 2016.

N. Shahbazi, Y. Lin, A. Asudeh, and H. Jagadish. Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*, 55(13s):1–39, 2023.

B. Shi, J.-t. Huang, G. Li, X. Zhang, and Z. Yao. Fairgamer: Evaluating biases in the application of large language models to video games. *arXiv preprint arXiv:2508.17825*, 2025.

H.-C. Shin, Y. Peng, E. Bodur, and et al. Biomegatron: Larger biomedical domain language model. In *Proceedings of EMNLP*, 2020. URL https://aclanthology.org/2020.emnlp-main.379/.

N. Singh, K. Lawrence, S. Richardson, and D. M. Mann. Centering health equity in large language model deployment. *PLOS Digital Health*, 2(10):e0000367, 2023.

K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023a.

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023b.

SKMCH&RC. Shaukat khanum memorial cancer hospital & research centre cancer registry. Website, 2024. URL https://shaukatkhanum.org.pk/health-care-professionals-researchers/cancer-statistics/skm-cancer-registry/.

I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. W. Kim, S. Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

A. Swaminathan, S. Salvi, P. Chung, A. Callahan, S. Bedi, A. Unell, M. Kashyap, R. Daneshjou, N. Shah, and D. Dash. Feasibility of automatically detecting practice of race-based medicine by large language models. In *AAAI 2024 spring symposium on clinical foundation models*, 2024.

T. Templin, S. Fort, P. Padmanabham, P. Seshadri, R. Rimal, J. Oliva, K. Hassmiller Lich, S. Sylvia, and N. Sinnott-Armstrong. Framework for bias evaluation in large language models in healthcare settings. *npj Digital Medicine*, 8(1):414, 2025.

Y. Tian, M. Shi, Y. Luo, A. Kouhana, T. Elze, and M. Wang. Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In *The Twelfth International Conference on Learning Representations*, 2023.

Y. Tian, Y. Luo, F. Liu, et al. Fairseg: A large-scale medical image segmentation dataset for fairness learning. *ICLR 2024 (proc. abstract) / arXiv:2311.02189*, 2024. URL https://ophai.hms.harvard.edu/datasets/harvard-fairseg10k.

A. A. Tierney, M. E. Reed, R. W. Grant, F. X. Doo, D. D. Payán, and V. X. Liu. Health equity in the era of large language? models. *American Journal of Managed Care*, 31(3), 2025.

E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6 (12):1399–1406, 2022.

E. K. Tokpo and T. Calders. Text style transfer for bias mitigation using masked language modeling. *arXiv preprint arXiv:2201.08643*, 2022.

P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 2018. URL https://www.nature.com/articles/sdata2018161.

I. Turcan and K. McKeown. Dreaddit: A reddit dataset for stress analysis in social media. In *EMNLP-IJCNLP*, 2019. URL https://arxiv.org/abs/1911.00133.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

K. Vishnubhotla, P. Cook, and G. Hirst. Emotion word usage in tweets from us and canada (tusc). *arXiv preprint arXiv:2204.04862*, 2022. URL https://arxiv.org/abs/2204.04862.

D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, X. Gao, J. Shen, J. He, T. Shen, Q. Duan, J. Zhao, K. Li, Y. Qiao, and S. Zhang. ChestDR: Thoracic Diseases Screening in Chest Radiography, 8 2023a. URL https://springernature.figshare.com/articles/dataset/ChestDR_Thoracic_Diseases_Screening_in_Chest_Radiography/22302775.

D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, X. Gao, J. Shen, J. He, T. Shen, Q. Duan, J. Zhao, K. Li, Y. Qiao, and S. Zhang. ColonPath: Tumor Tissue Screening in Pathology Patches, 8 2023b. URL https://springernature.figshare.com/articles/dataset/ColonPath_Tumor_Tissue_Screening_in_Pathology_Patches/22302799.

D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, X. Gao, J. Shen, J. He, T. Shen, et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Scientific Data*, 10(1):574, 2023c.

D. Wang, X. Wang, L. Wang, M. Li, Q. Da, X. Liu, et al. Neojaundice: Neonatal jaundice evaluation in demographic images. https://springernature.figshare.com/articles/dataset/NeoJaundice_Neonatal_Jaundice_Evaluation_in_Demographic_Images/22302559, 2023d.

H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, and T. Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023e. URL https://arxiv.org/abs/2304.06975.

H. Wang, S. Zhao, Z. Qiang, N. Xi, B. Qin, and T. Liu. Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis. *arXiv preprint arXiv:2401.16107*, 2024a.

W. Wang, H. Bai, J.-t. Huang, Y. Wan, Y. Yuan, H. Qiu, N. Peng, and M. Lyu. New job, new gender? measuring the social bias in image generation models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3781–3789, 2024b.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017. URL https://arxiv.org/abs/1705.02315.

Y. Wang, Y. Zhao, S. A. Keller, A. de Hond, M. M. van Buchem, M. Pillai, and T. Hernandez-Boussard. Unveiling and mitigating bias in mental health analysis with large language models. *arXiv preprint arXiv:2406.12033*, 2024c.

Y. Wang et al. Unveiling and mitigating bias in mental health analysis using large language models. *arXiv preprint arXiv:2406.12033*, 2024d. URL https://arxiv.org/abs/2406.12033. Uses the DepEmail dataset.

Z. Wang, Z. Wu, D. Agarwal, and J. Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of EMNLP*, 2022. URL https://aclanthology.org/2022.emnlp-main.256.pdf.

J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

Writer. Palmyra med: Instruction-based fine-tuning of llms enhancing medical domain performance. https://writer.com/engineering/palmyra-med-instruction-based-fine-tuning-medical-domain-performance/, 2023.

C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024a.

J. T. Wu, M. Moradi, H. Wang, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021. URL https://physionet.org/content/chest-imagenome/1.0.0/.

P. Wu, C. Liu, C. Chen, J. Li, C. I. Bercea, and R. Arcucci. Fmbench: Benchmarking fairness in multimodal large language models on medical tasks. *arXiv preprint arXiv:2410.01089*, 2024b.

xAI. Grok 4 model card. Technical report, xAI, Aug. 2025. URL https://x.ai/news/grok-4.

Y. Xiao, J. M. Zhang, Y. Liu, M. R. Mousavi, S. Liu, and D. Xue. Mirrorfair: Fixing fairness bugs in machine learning software via counterfactual predictions. *Proceedings of the ACM on Software Engineering*, 1(FSE):2121–2143, 2024.

Y. Xiao, J. Huang, R. He, J. Xiao, M. R. Mousavi, Y. Liu, K. Li, Z. Chen, and J. M. Zhang. Amqa: An adversarial dataset for benchmarking bias of llms in medicine and healthcare. *arXiv preprint arXiv:2505.19562*, 2025.

H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, L. Huang, Q. Wang, and D. Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*, 2023. URL https://arxiv.org/abs/2304.01097.

J. Xu, D. Ju, M. Li, Y.-L. Boureau, J. Weston, and E. Dinan. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*, 2020.

B. Yan, W. Zeng, Y. Sun, W. Tan, X. Zhou, and C. Ma. The guideline for building fair multimodal medical ai with large vision-language model, 2024.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

K. Yang, C. Yu, Y. R. Fung, M. Li, and H. Ji. Adept: A debiasing prompt framework. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 10780–10788, 2023.

X. Yang, Y. Wu, and et al. A large language model for electronic health records. *npj Digital Medicine*, 2022. URL https://www.nature.com/articles/s41746-022-00742-2.

Y. Yang, X. Liu, Q. Jin, F. Huang, and Z. Lu. Unmasking and quantifying racial bias of large language models in medical report generation. *ArXiv*, 2024.

Y. Yang, Y. Liu, X. Liu, A. Gulhane, D. Mastrodicasa, W. Wu, E. J. Wang, D. Sahani, and S. Patel. Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances*, 11(13):eadq0305, 2025b.

J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*, 2023.

Y. H. Yeo, Y. Peng, M. Mehra, J. Samaan, J. Hakimian, A. Clark, K. Suchak, Z. Krut, T. Andersson, S. Persky, et al. Evaluating for evidence of sociodemographic bias in conversational ai for mental health support. *Cyberpsychology, Behavior, and Social Networking*, 28(1):44–51, 2025.

C. C. Young, E. Enichen, A. Rao, and M. D. Succi. Racial, ethnic, and sex bias in large language model opioid recommendations for pain management. *Pain*, pages 10–1097, 2022.

T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Xue, S. Vyas, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *EMNLP*, 2018. URL https://arxiv.org/abs/1809.08887.

T. Zack, E. Lehman, M. Suzgun, J. A. Rodriguez, L. A. Celi, J. Gichoya, D. Jurafsky, P. Szolovits, D. W. Bates, R.-E. E. Abdulnour, et al. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22, 2024.

P. S. Zahraei and Z. Shakeri. Detecting bias and enhancing diagnostic accuracy in large language models for healthcare. *arXiv preprint arXiv:2410.06566*, 2024.

A. Zhang, M. Yuksekgonul, J. Guild, J. Zou, and J. Wu. Chatgpt exhibits gender and racial biases in acute coronary syndrome management. *medRxiv*, pages 2023–11, 2023a.

J. M. Zhang and M. Harman. " ignorance and prejudice" in software fairness. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 1436–1447. IEEE, 2021.

K. Zhang and D. Liu. Customized segment anything model for medical image segmentation (samed). *arXiv preprint arXiv:2304.13785*, 2023. URL https://arxiv.org/abs/2304.13785.

S. Zhang, Z. Wang, and et al. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023b. URL https://arxiv.org/abs/2303.00915.

W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023a.

Z. Zhao, Q. Jin, F. Chen, T. Peng, and S. Yu. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, 10(909), 2023b. URL https://github.com/pmc-patients/pmc-patients.

G. Zheng, M. A. Jacobs, V. Braverman, and V. S. Parekh. Towards fair medical ai: Adversarial debiasing of 3d ct foundation embeddings. *arXiv preprint arXiv:2502.04386*, 2025a.

Y. Zheng, W. Gan, Z. Chen, Z. Qi, Q. Liang, and P. S. Yu. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*, 16(2):1015–1040, 2025b.