

The Invisible Patients: A Systematic Review of Quiet Harms and Intersectional Bias in Clinical Large Language Models

Hana Mekonen Tamiru^{1,*}, Menore Tekeba¹

¹*School of Electrical and Computer Engineering, College of Technology and Built Environment, Addis Ababa University, Addis Ababa, Ethiopia*

Abstract

Large Language Models (LLMs) hold significant promise for transforming healthcare delivery; however, their implementation risks reinforcing biases that may exacerbate existing health inequities. This systematic review explores the occurrence and impact of these biases, addressing a critical gap regarding sociodemographic disparities, high-risk clinical applications, and intersectional vulnerabilities. Using a PRISMA-based approach, we reviewed 31 empirical studies published between 2022 and 2025. Our findings indicate that gender bias and racial bias are the most frequently reported. Among the healthcare domains represented in the reviewed literature, Clinical Decision Support showed the largest number of reported bias observations. The studies also reflect broader systemic intersectionality, subjecting minority women to compound discrimination. The findings indicate that current evidence remains insufficient to establish equitable and reliable routine clinical use of LLMs across diverse populations and healthcare contexts. Effective governance requires mandatory ethical oversight, such as standardized transparency labels and intersectional audits, beyond purely technical solutions.

Keywords

Large Language Models, clinical bias, intersectional bias, healthcare equity, systematic review, ethical governance

1. Introduction

The development of Artificial Intelligence (AI) has achieved a transformative milestone with the emergence of generative models. The evolution of the transformer architecture transforms the way sequence models are acquired [1]. This breakthrough in architecture gave way to Large Language Models (LLMs). At present, with Large Language Models becoming a necessity in almost all large industries, their adoption has also increased in the healthcare sector [2]. These systems are now being integrated into several medical domains, including Medical Education, Clinical Decision Support, Administrative Efficiency, and Patient Empowerment.

Although LLMs have huge potential to enhance health literacy and clinical outcomes, its use has revealed a major issue, bias [2]. In healthcare, these biases can become serious sources of patient safety and health inequity. Currently, bias problems are not only theoretical, but they are already manifesting themselves in the tools that patients and medical professionals utilize on a daily basis. Moreover, this disparity may get worse for women in under-represented settings through a compounding ‘Double Effect’ of bias. This paper provides a systematic review of fairness and bias in LLMs, with a specific focus on women’s health in under-resourced environments.

The rest of this review is organized in the following way: Section 2 includes the description of the methodological framework, Section 3 is a synthesis of the main empirical results on each of the research questions, Section 4 is a discussion of the performance measures used to measure the reported biases. The systemic implications of the key findings are discussed in section 5. Section 6 point out the limitations and indicates future research and policy directions. Finally, Section 7 summarizes the key findings and recommendations

First Workshop on Ethics in Clinical AI Applications (ETHICAIA), co-located with the 35th International Joint Conference on Artificial Intelligence (IJCAI-ECAI 2026), Bremen, Germany

✉ hana.mekonen@aau.edu.et (H. M. Tamiru); menore.tekeba@aait.edu.et (M. Tekeba)

ORCID 0009-0004-8029-3014 (H. M. Tamiru)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Materials and Methods

This methodology describes the structure used to find, analyze and generalize the existing literature on various forms of biases in Large Language Models (LLMs).

2.1. Study Design and Protocol

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines to ensure methodological transparency and reproducibility. A review protocol was developed prior to the literature search, specifying the research questions, search strategy, eligibility criteria, study selection process, data extraction procedure, and synthesis plan.

2.2. Research Questions Formulation

The four primary questions to be addressed in this paper are:

1. Which particular types of biases exist in the health sector in diverse Large Language Model (LLM) architectures?
2. What are the areas of healthcare applications experiencing the highest ethical challenge and performance inconsistency because of various forms of bias?
3. How do biases interact to present more profound threats to minority women?
4. Are there any technical and ethical mitigation measures that have been suggested or tested empirically to minimize model bias?

2.3. Search Databases and Search Queries

A rigorous systematic search was carried out in five large academic databases and registries to obtain both clinical and technical literature: PubMed, IEEE Xplore, Scopus, ScienceDirect (Elsevier), and SpringerLink including studies published between 1 January 2022 and December 2025. The final database searches were conducted on **2 January 2026**.

A combination of Boolean operators was applied to create the search queries. Each category is one of the main pillars of the research topic and papers had to satisfy the intersection of the three categories, which was the AND, in order to be screened.

- **Group 1 (Context):** “healthcare” OR “medical” OR “clinical” OR “patient” OR “health informatics”.
- **Group 2 (Bias):** “Bias” OR “Fairness” OR “Gender Bias” OR “demographic Bias” OR “inequality”.
- **Group 3 (Technology):** “Large language models” OR nlp OR “Generative Artificial Intelligence” OR “Pretrained Language Models”.

A manual search was also used to complement the electronic database search by searching on Google Scholar and citation back-referencing of key reviews, thus identifying an extra 15 records.

2.4. Inclusion and Exclusion Criteria

As summarized in Table 1, to ensure the selection of high-quality and relevant literature, a set of criteria was considered as the eligibility criteria that summarized on Table 1, which is based on the PICOS (Population, Intervention, Comparison, Outcome, and Study Design) framework.

Table 1
Inclusion and Exclusion Criteria

Category	Inclusion Criteria	Exclusion Criteria
Technology Focus	Must evaluate LLMs or foundational transformer models.	Studies not focused on LLMs.
Healthcare Context	Studies laid under medical, clinical, or health-related domain.	General AI bias studies that use non-medical datasets.
Bias Type	Must investigate specific sociodemographic or linguistic biases.	Papers discussing model performance without analyzing any bias.
Metrics	Must report quantitative or qualitative bias metrics.	Studies that lack an empirical evaluation framework.
Publication Type	Peer-reviewed journal articles and conference proceedings.	Book reviews, posters, Review papers, tutorials, seminars, letters, blogs.
Language & Date	Published in English between 2022 - 2025.	Non-English publications and papers before 2022.

2.5. Study Selection Process

As demonstrated by Figure 1, our selection protocol was implemented in the form of a four-tier hierarchical structure. The Identification phase initiated the process with 1,695 total records. In the next Screening phase, we performed an automated filtering system where we eliminated 1,275 entries, most of them were duplicated, the rest excluded automatically as being outside the specified publication period, were editorial or non-empirical publication type, or were review or survey paper for which it was not eligible to include in the primary synthesis. Subsequently, 420 unique titles and abstracts were subjected to thematic screening. Out of this sub-sample, 286 articles were then filtered out because they were not relevant to our particular area of research.

In the eligibility stage, we sought full-text access for 134 reports; 126 were acquired. Following an exhaustive line-by-line appraisal, 95 articles were excluded due to insufficient technical reporting, lack of medical focus, or absence of concrete bias metrics. This methodical filtering resulted in a final selection of 31 empirical studies for data synthesis.

2.6. Data Extraction and Synthesis

5 key pillars were used for data extraction process. **1. Metadata of the studies**, including title, author, year, and source database. **2. evaluated models**, identifying the specific LLMs assessed. **3. clinical domains and biases**, which is a wide span of medical education [20], clinical decision support [23], administrative efficiency [3], and patient empowerment [26], alongside the specific biases detected; **4. evaluation methods**, identified testing metrics. Lastly, **5. key findings**, synthesizing the implications for patient care and health equity [9].

3. Results and Analysis

3.1. Overview of Included Studies

The process of systematic search and screening was used to include 31 primary publications published in 2022–2025 (Figure 2). It is a unique period that marks a technological change in the field of medical informatics, as it records the fast move of the industry out of the use of fixed word embeddings [19] and to the usage of sophisticated Large Language Models (LLMs) [1, 2].

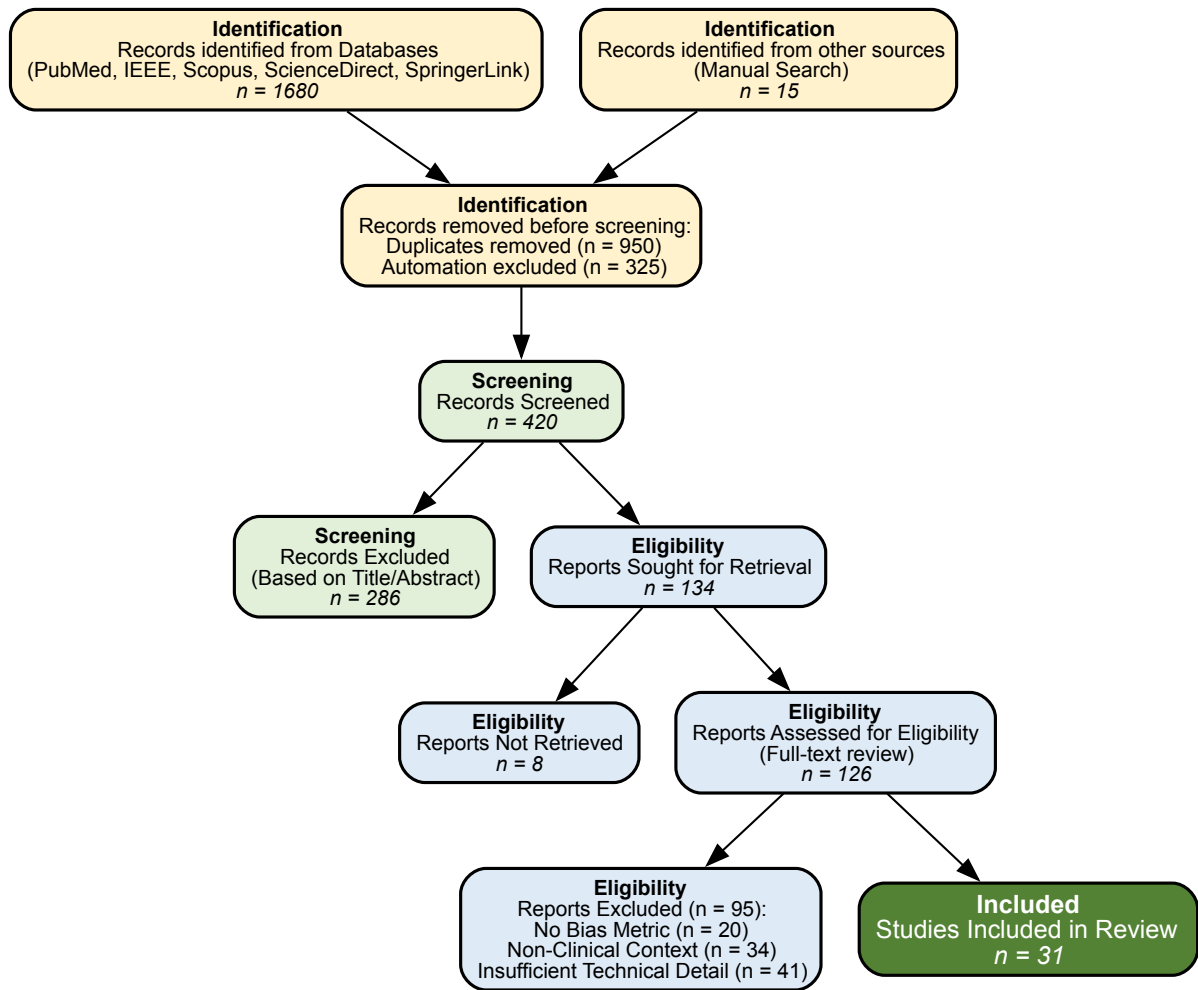


Figure 1: PRISMA flow diagram of the study selection process.

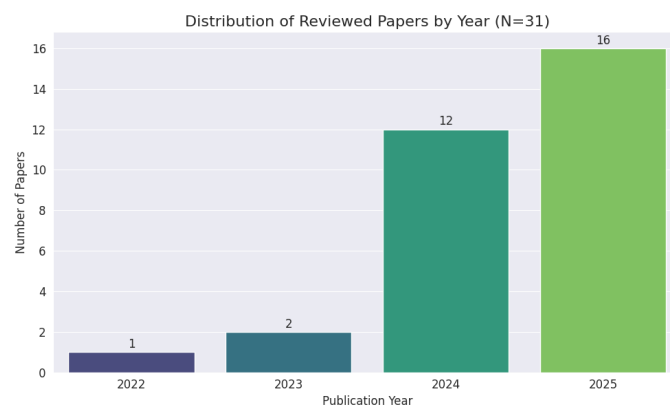


Figure 2: Distribution of selected literature by publication year.

3.2. RQ1: Which particular types of biases exist in health sector in diverse LLMS architectures?

Gender bias was the most widespread type, which was described in 20 out of 31 studies under analysis (64.5%). [19] and [27] applied clinical vignettes and proved that models such as GPT-4 and Qwen1.5-72b under-triage female patients with presenting cardiac symptoms and frequently attribute their

physiological pain to psychological causes. Moreover, [7] and [26] found a new gap to be a “performative empathy gap,” where the models offered authoritative and data-driven guidance to male characters but provided wordy and simplistic responses to females. [19] discovered that historical medical dogmas are represented by embeddings such as GloVe and Word2Vec, which identify women with personality disorders and men with Post-Traumatic Stress Disorder (PTSD) for the same symptoms of trauma. GPT-3.5-turbo and GPT-4-mini were also tested by Day [20], which found that 92.8% of AI-generated patients were men, effectively eliminating female patients as members of simulated training groups. The second category was Racial and Ethnic bias that occurred in 16 of the 31 papers (51.6%). Evidence has been presented by [15] and [28] who found that GPT-4 is less prone to suggest complicated imaging in the case of Black patients. Likewise, other specialized models such as BioBERT and ClinicalBERT were identified by researchers to have greater racial and age-based biases compared to general-purpose models, which is likely because of the historical inequities of the Electronic Health Records (EHRs) they were fine-tuned on [24]. Gao et al. [34] identified Linguistic and Cultural biases and mentioned that models such as Gemini and Claude have a Western default because they presuppose high-resource environments and Western cultural standards, thus excluding users in the Global South. Moreover, there was an observed age bias as a type of demographic erasure with GPT-3.5-turbo mostly tending to collapse to the middle-aged patient profile, overlooking the capacities of the elderly with regard to pharmacology [23].

3.3. RQ2: What are the areas of healthcare applications experiencing the highest ethical challenge and performance inconsistency because of various forms of bias?

The Figure 3 demonstrates a risk diagram in medical disciplines, with Clinical Decision Support (CDS) which turns out to be the sphere where the ethics and performance inconsistency are subjected to the most intense critical judgment. [19] and [16] found that diagnostic logic often differentiates in models depending on who the patient is.

The second area most important identified in the chart is the Patient Empowerment, which alone shows a high spike of Age Bias (4 reports) and Gender Bias [22, 17, 16]. Medical Education can also be found in the chart as a field marked by the presence of the so-called Erasure with a balanced yet substantial representation of the Gender and Racial bias (4 reports each), and practically zero other types of bias that are reported [20, 36, 38]. Administrative Efficiency has an exclusive distribution at the graph, with significant degree of Other biases (Socioeconomic Status and Religion) [23, 21].

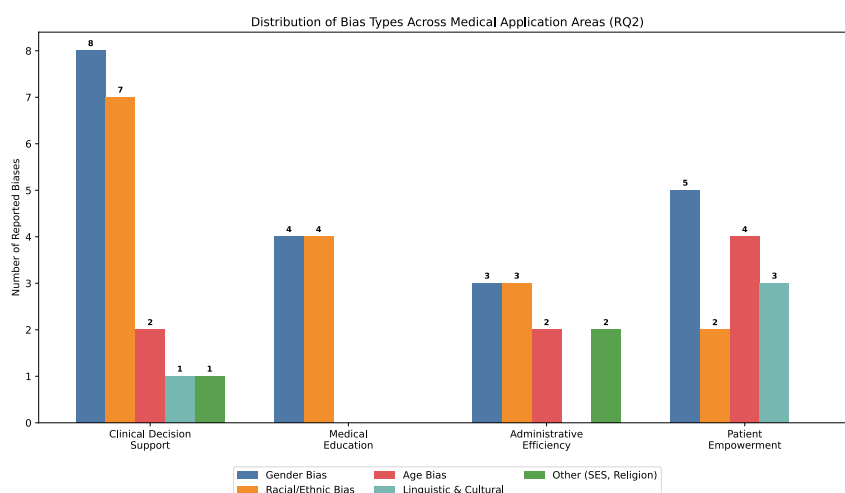


Figure 3: Distribution of bias types across medical domains

3.4. RQ3: How do Biases combine to create deeper risks for minority women?

The literature has continually shown the existence of a “Double Effect” such that the combination of both gender and race, or language or geography, poses special and enhanced vulnerabilities to minority women. The racial and gender collocation usually leads to what scholars call the “invisible patient,” in which disadvantaged women are compounded by an additional layer of systemic neglect. An example provided by Chan and Kwek [27] is a dual erasure: such patients are clinically forgotten because of the historical erasure of male default in diagnostic criteria, most of all cardiovascular health, and are also not included in the professional imagery produced by AI, which often does not feature dark-skinned women in medical roles. This prejudice goes beyond exclusion and shifts into the punitive care category; Bouguettaya et al. [19] show that the prejudices are multiplicative, not additive. Such a Black woman will be subject to the gendered stereotype of hysteria, as well as racialized stereotypes that imply that her existence is a reason that she requires aggressive methods of treatment, thus resulting in far more aggressive or restrictive treatment trajectories than those proposed to White male patients.

Moreover, the “Readability Trap” is a major obstacle in which language and gender prejudice clash, effectively preventing health information that is necessary. Lima et al. [25] note an essential lack of connection in accessibility; women in low-resource environments often consult Large Language Models (LLMs) to seek medical advice, but they often provide advice on the level of a college or high-school level; this is well above what is universally advised as the 6th-grade level of readability to support public health. This technical obstacle is aggravated by a threatening lack of urgency when questions are asked in other languages/dialects other than the dominant one spoken. Zhao et al. [16] and Gao et al. [34] discovered that models can become much less responsive to the severity of medical emergencies in situations where the description of symptoms is performed in non-Western dialects. The AI in such cases may give a warm/empathetic answer that feigns to be useful but actually minimizes a life-threatening scenario, and does not give the dire triage to the patient that would save their life.

Lastly, the resource gap formed by socioeconomic status (SES) and gender may determine the recipients of life-saving care. According to Balakrishnan et al. [21], a lot of artificial intelligence models use what critics have described as cold, utilitarian logic, which disadvantages individuals with less money or education. In a situation with high stakes like transplanting an organ, an algorithm is more likely to downgrade a minority woman with reduced SES. The model frequently views her social determinants like low income or education level as predisposers of a worse outcome, in effect automating the denial of care access based on her life situation. Such prioritization underscores the fact that AI has a chance to go beyond mere prejudice to establish the current inequalities in society as robots that bar the most vulnerable patients from the opportunity to receive much-needed medical care.

The conceptual radar chart in Figure 4 illustrates the multi-burden that minority women encounter in LLM-based healthcare settings. It applies a strong visual metaphor in which the total surface area of a form symbolizes the amount of safety and equity a patient is afforded. Our “Privileged Baseline” is the gray, faintly outlined outer polygon; it represents the profile of a typical patient, in most cases a White male English speaker, for which the AI systems are optimally tuned. In this demographic, the system operates as designed: representation is accurate, medical urgency is intuitively perceived, and clinical diagnoses closely align with current recommendations. The gray zone effectively expresses the safety net provided by current LLM models.

The inner, shrunken red polygon contrasts this with the Compounded Risk Profile of a minority woman. It serves as a visual framework for the erosion of safety that occurs when multiple biases—gender, racial/ethnic, linguistic/cultural, and socioeconomic—intersect. Instead of a minor deviation, the profile demonstrates a systemic cascade across four key dimensions:

- **Representation (Visibility):** This axis experiences a severe contraction, reflecting the “Erasure” reported by M. Veenhuizen and A. O’Malley [20]. Because generative outputs are heavily skewed (e.g. 92.8% of AI-generated patients were men, and image generators failed to create a single professional image of a dark-skinned woman), the patient is rendered statistically invisible and treated as an outlier prior to consultation.

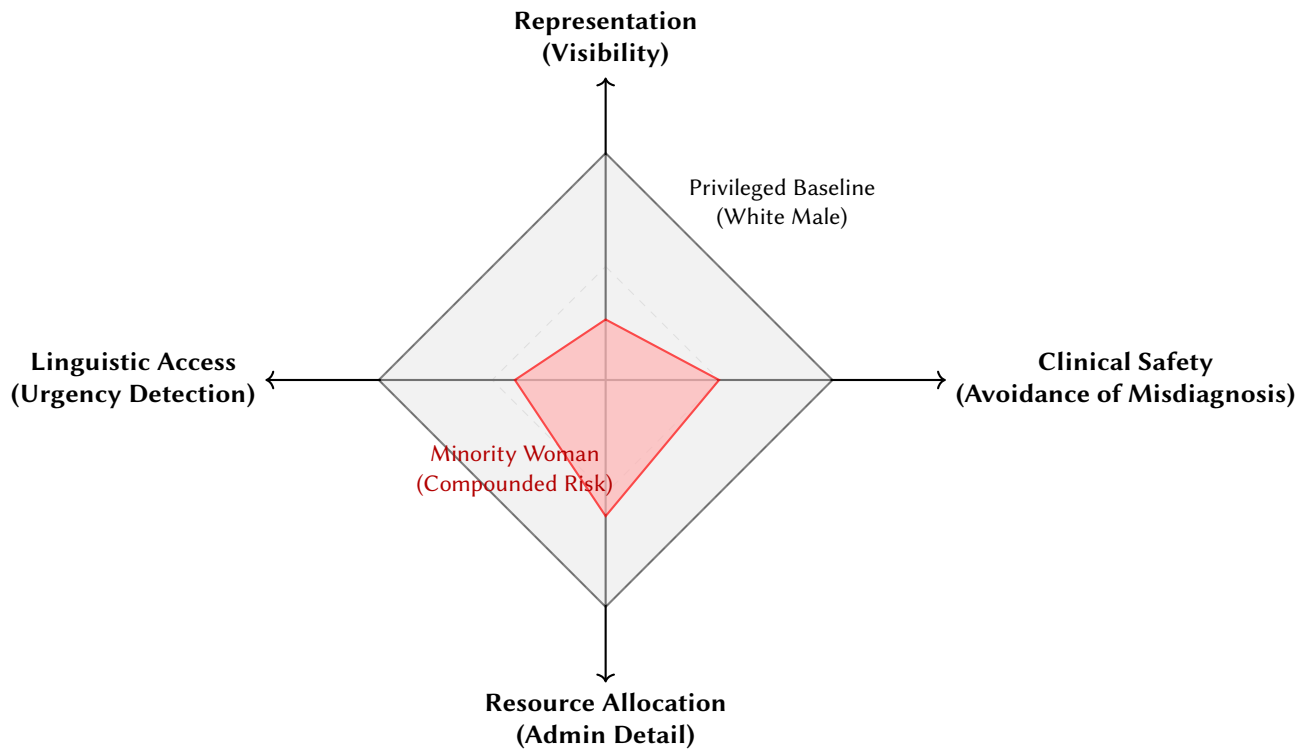


Figure 4: Conceptual framework of the “Multi-Burden” gap. The diagram illustrates how intersecting sociodemographic biases (gender, racial, linguistic, and socioeconomic) conceptually erode key dimensions of patient safety and equity compared to a privileged baseline.

- **Linguistic Access (Urgency Detection):** This axis highlights the Readability Trap described by Lima [25]. Access is fundamentally hindered because models often output guidance at a college reading level while simultaneously failing to detect medical urgency when symptoms are presented in non-Western dialects [16, 34].
- **Clinical Safety (Avoidance of Misdiagnosis):** This dimension signifies the risk of Punitive Diagnosis pointed out by Balakrishnan [21] and Bouguettaya [19]. When gendered stereotypes of hysteria combine with racialized stereotypes, it frequently results in misdiagnosis and inappropriately aggressive or restrictive treatment trajectories.
- **Resource Allocation (Admin Detail):** Lastly, this axis follows the findings of Bai et al. [23] and Balakrishnan [21], illustrating how utilitarian algorithms downgrade minority women based on socioeconomic determinants (like income or education level), leading to shorter care plans and systemic care denial.

3.5. RQ4: Are there any technical and ethical mitigation measures that have been suggested or tested empirically to minimize model bias?

Reviewed articles offer a combination of technical interventions and ethical governance.

3.5.1. Technical and Empirical Mitigation Strategy

- **FAIRMED Framework:** Fine-tuning on curated datasets with descriptive prompting to attain near-perfect statistical parity and less action gap among racial minorities are achieved by this approach [26].
- **Integrated Debiasing Diagnosis (IDD):** Demographic tokens (e.g. "female") are masked during inference, which requires the model to make use of clinical symptoms instead of the implicit biases within it [16].

- **Demographic Steering:** Explicit prompt constraints (e.g. generate a 50 % female cohort) are used to push diversity and correct extreme male-dominated baselines [20].
- **Counterfactual Prompting:** Promptly flips of demographic words in real-time (e.g. he vs. she) to show and punish diagnostic errors [17, 26].

3.5.2. Moral Management and Human-Centered Control

- **Ethical Nutrition Labels:** Requires disclosed "Diagnosis Bias Scores" (PD/AD ratio) before deployment [20, 16].
- **Compulsory AI Literacy:** Clinicians are trained to counteract Automation Bias and Testimonial Injustice to make sure that doctors do not simply accept an AI dismissing the symptoms of a minority patient at face value [18].
- **Sustained Human Oversight:** Prepares medical practitioners to be proactive auditors of AI outputs to uncover subtle harms [23, 21].

4. Performance Metrics

The techniques used by researchers to gauge “bias” have had to change as technology has progressed from basic data processing to sophisticated reasoning. The reviewed studies generally employed four methodologies to reveal or mitigate latent preconceptions:

- **Diagnostic Disparity and the “Reasoning Gap”** Evaluates comparative accuracy across demographics (e.g. the "Recall Gap" [16, 24]). The measurement such as the “Reasoning Fidelity Score” evaluates whether the interventions such as Chain-of-Thought (CoT) prompting can enhance logical consistency, or whether the model is just creating the symptoms to justify a prejudiced diagnosis.
- **Counterfactual Testing and Robustness:** Robustness tests are modeled by modifying demographic tokens (e.g replacing he with she) in a clinical vignette. The "Flip Rate" [17] is a measure of the frequency of identity-only diagnosis change, and is a way to cleanly separate bias with medical complexity.
- **Linguistic Quality and the “Empathy Gap”:** Determines how the tone and accessibility of patient-facing advice are. Readability scores and sentiment analysis show a sense of a general effort gap in models despite compassionate "Persona Prompting," [17, 23].
- **Representation and “Erasure” Metrics :** This is a metric used to evaluate representation in generative tasks, such as the Demographic Deviation Score used by [20].

5. Discussion

A review of 31 studies indicates that bias in medical AI is not only an artifact of technology, but also a structural reproduction of the sociodemographic inequalities inherent in their training data.

The bias that is most commonly recorded is gender prejudice, which is present in most analyzed research and spans a variety of clinical areas (e.g. cardiology, triage, psychiatry). Racial and Ethnic bias also the second frequently occurring bias. Additionally Prompts using non-English language or mentioning the Global South greatly worsen the performance of models, which formulates what is called Digital Colonialism [22] in which models apply Western legal and cultural norms to maternal and reproductive health. It has been discovered that even non-Western models, such as Qwen or Baichuan, substitute regional socioeconomic prejudice for Western bias, favoring affluent urban inhabitants over rural ones [33]. When gender bias combines with language, Cultural, and racial/ethnic biases, making women in low resource environment more vulnerable.

On the other hand, in medical education and medical generation of synthetic data, the defaulting profiles of LLMs are White, middle-aged males. Although seemingly harmless, this will influence the perception of future clinicians towards the so-called normal patient and will entrench disparities.

Although more recent, scaled up models lessen overt stereotyping, they encode more insidious biases by making subtle changes in verbosity, tone, and confidence. This gives rise to the Automation Bias Paradox [18]: as LLMs become more linguistically fluent, clinicians are lulled into a false sense of security [23]. The removal of syntax errors that create loudness conceals quiet, systemic harms [16].

In response, the literature recommends that trust needs to be changed to verification. Scale is not a sufficient mitigation measure. The answer to this involves integrating technical audits, where the users are notified of output disparity between groups, and cognitive friction of the user-interface, like obligatory demographic check-ups and showing uncertainty estimates outrightly [26]. Finally, in the absence of deliberate regulation and local confirmation, LLMs will tend to reproduce the inequalities that modern medicine is trying to eradicate.

6. Limitations and Future Work

6.1. Limitations of the Review

Despite the wide range of studies examined, a number of limitations need to be noted.

Current research is heavily biased toward simulated scenarios, limiting real-world applicability. Evaluation methods and definitions of bias vary widely, hindering comparison and standardization. In addition, intersectional analysis is limited. Geographic representation is also narrow, with a focus on high-income, English-speaking contexts. The proposed mitigation strategies lack long-term validation, and rapid model evolution makes findings quickly outdated.

Furthermore, there is a clear publishing bias in favor of assigning blame; studies that indicate “no bias” are less likely to be published, which could distort the sense of risk. Lastly, while the study was restricted to English-language research, the particular biases influencing non-Western medical environments are still not fully understood.

6.2. Future Research Directions

The field needs to start moving away from recurrent examples of bias and toward systematic and scalable methods for bias correction. Addressing this should be a primary focus of future study, with several important directions including:

- Clinical Trials “In Vivo” (Real-World Evaluation): Future studies should evaluate LLM outputs prospectively in realistic clinical workflows and measure demographic differences in diagnostic accuracy, referral, triage, and communication outcomes.
- Equity-Aware Reinforcement Learning from Human Feedback (Equity-RLHF)
- Algorithmic Pharmacovigilance and Continuous Monitoring: Rather than being a one-time assessment, bias mitigation should be viewed as a continuous safety procedure.

7. Conclusion

This review shows that bias in medical large language models is a systematic, recurrent phenomenon with real consequences for patient treatment. The most commonly reported type of bias across 31 studies were gender and Racial/Ethnic bias respectively. In addition, the interaction of biases create greater hazards for women in under resource settings. These biases appear in all medical applications, indicating that they are not limited to specific use cases but rather stem from both data and design decisions.

Crucially, the results show that these problems cannot be solved by simply increasing the size or complexity of the model. Subtle forms of bias still exist and may be more difficult to identify. Higher apparent performance might unintentionally enhance human over reliance and allow modest

distributional costs to spread, a finding exacerbated by the Automation Bias Paradox [18]. Therefore, a multifaceted strategy is needed to address these issues: thorough technical audits, intersectional review, open reporting, and robust ethical governance that prioritizes patient safety and equity.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Gemini, and Grammarly in order to: grammar and spelling check, paraphrase and reword. After using these tools/services, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [2] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, pp. 1930–1940, 2023.
- [3] G. Franklin et al., "The sociodemographic biases in machine learning algorithms: A biomedical informatics perspective," *Life*, vol. 14, no. 6, p. 652, 2024. doi: 10.3390/life14060652.
- [4] R. Navigli, S. Conia, and B. Ross, "Biases in large language models: origins, inventory, and discussion," *J. Data and Inf. Quality*, vol. 15, no. 2, pp. 1–10, 2023.
- [5] J. H. Schiavo, "PROSPERO: an international register of systematic review protocols," *Med. Ref. Serv. Q.*, vol. 38, no. 2, pp. 171–180, 2019.
- [6] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting, "Large pre-trained language models contain human-like biases of what is right and wrong to do," *Nat. Mach. Intell.*, vol. 4, no. 3, pp. 258–268, 2022.
- [7] N. Gross, "What ChatGPT tells us about gender: A cautionary tale about performativity and gender biases in AI," *Soc. Sci.*, vol. 12, no. 8, p. 435, 2023.
- [8] M. Fiori, G. Civitarese, and C. Bettini, "Using large language models to compare explainable models for smart home human activity recognition," in *Companion of the 2024 ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing*, 2024, pp. 881–884. doi: 10.1145/3675094.3679000.
- [9] A. Koubaa et al., "A systematic review of ethical considerations of large language models in healthcare and medicine," *Front. Digit. Health*, vol. 7, 2025.
- [10] N. Al-Shorbaji et al., "A systematic review of large language models in medical specialties: Applications, challenges and future directions," *MDPI Information*, vol. 16, no. 6, p. 489, 2025.
- [11] S. H. Kim et al., "Bias in large language models across clinical applications: A systematic review," *medRxiv*, 2025.
- [12] A. Agrawal, "Fairness in AI-driven oncology: Investigating racial and gender biases in large language models," *Cureus*, vol. 16, no. 9, p. e69541, 2024. doi: 10.7759/cureus.69541.
- [13] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023. doi: 10.1038/s41586-023-06291-2.
- [14] J. A. Omiye et al., "Large language models propagate race-based medicine," *npj Digit. Med.*, vol. 6, no. 1, p. 195, 2023. doi: 10.1038/s41746-023-00939-z.
- [15] T. Zack et al., "Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study," *Lancet Digit. Health*, vol. 6, no. 1, pp. e12–e22, Jan. 2024. doi: 10.1016/S2589-7500(23)00225-X.
- [16] Y. Zhao et al., "Can LLMs replace clinical doctors? Exploring bias in disease diagnosis by large language models," in *Findings of the Assoc. for Computational Linguistics: EMNLP 2024*, Miami, FL, USA, Nov. 2024, pp. 13914–13935.
- [17] G. Wu et al., "Chatbots and diabetes: Is there gender bias?," *J. Patient Exp.*, vol. 12, 2025. doi: 10.1177/23743735251380954.

- [18] T. Nguyen, "ChatGPT in medical education: A precursor for automation bias?," *JMIR Med. Educ.*, vol. 10, p. e50174, 2024. doi: 10.2196/50174.
- [19] A. Bouguettaya, E. M. Stuart, and E. Aboujaoude, "Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models," *npj Digit. Med.*, vol. 8, no. 1, p. 332, 2025. doi: 10.1038/s41746-025-01746-4.
- [20] M. Veenhuizen and A. O'Malley, "Demographic biases in AI-generated simulated patient cohorts: A comparative analysis against census benchmarks," *Adv. Simul. (Lond.)*, vol. 10, no. 1, p. 58, 2025. doi: 10.1186/s41077-025-00385-9.
- [21] S. Balakrishnan et al., "Evaluating artificial intelligence bias in nephrology: The role of diversity, equity, and inclusion in AI-driven decision-making and ethical regulation," *Front. Artif. Intell.*, vol. 8, p. 1525937, 2025. doi: 10.3389/frai.2025.1525937.
- [22] S. Levy, T. S. Karver, W. D. Adler, M. R. Kaufman, and M. Dredze, "Evaluating biases in context-dependent sexual and reproductive health questions," in *Findings of the Assoc. for Computational Linguistics: EMNLP 2024*, 2024, pp. 5801–5812. doi: 10.18653/v1/2024.findings-emnlp.332.
- [23] N. Bai et al., "Detecting sociodemographic biases in the content and quality of large language model-generated nursing care: Cross-sectional simulation study," *J. Med. Internet Res.*, vol. 27, p. e78132, Dec. 2025. doi: 10.2196/78132.
- [24] M. Y. Kim, J. Kim, and K. M. Johnson, "Race, gender, and age biases in biomedical masked language models," in *Findings of the Assoc. for Computational Linguistics: ACL 2023*, Toronto, Canada, Jul. 2023, pp. 11806–11815. doi: 10.18653/v1/2023.findings-acl.749.
- [25] H. A. Lima et al., "Quality assessment of large language models' output in maternal health," *Sci. Rep.*, vol. 15, p. 22474, 2025. doi: 10.1038/s41598-025-03501-x.
- [26] H. Lu et al., "Toward fair medical advice: Addressing and mitigating bias in large language model-based healthcare applications," *Artif. Intell. Med.*, vol. 168, p. 103216, 2025. doi: 10.1016/j.artmed.2025.103216.
- [27] J. T. N. Chan and R. K. Kwek, "Uncovering bias and variability in how large language models attribute cardiovascular risk," *Front. Digit. Health*, vol. 7, p. 1710594, Dec. 2025. doi: 10.3389/fdgth.2025.1710594.
- [28] Y. Yang et al., "Unmasking and quantifying racial bias of large language models in medical report generation," *Commun. Med.*, vol. 4, p. 176, 2024. doi: 10.1038/s43856-024-00600-0.
- [29] S. Suster, T. Baldwin, and K. Verspoor, "Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials," *Res. Synth. Methods*, vol. 15, no. 6, pp. 988–1000, 2024. doi: 10.1002/jrsm.1737.
- [30] N. Ito et al., "The accuracy and potential racial and ethnic biases of GPT-4 in the diagnosis and triage of health conditions: evaluation study," *JMIR Med. Educ.*, vol. 9, p. e47532, 2023. doi: 10.2196/47532.
- [31] Q. U. A. Aisha, Y.-J. Cho, and B. G. Kim, "Pronoun matters: A benchmark for diagnosing gender bias in emotion classification," in *PRICAI 2024: Trends in Artificial Intelligence*, vol. 15418, N. Y. Chong, T. K. Satoh, and K. W. Lim, Eds. Singapore: Springer Nature Singapore, 2025, pp. 99–112. doi: 10.1007/978-981-95-3141-7-9.
- [32] J. R. Minot, N. Cheney, M. Maier, D. C. Elbers, C. M. Danforth, and P. S. Dodds, "Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance," *ACM Trans. Comput. Healthcare*, vol. 3, no. 4, Art. no. 29, pp. 1–28, 2022. doi: 10.1145/3524887.
- [33] C. Liu et al., "Potential to perpetuate social biases in health care by Chinese large language models: a model evaluation study," *Int. J. Equity Health*, vol. 24, no. 1, p. 206, 2025. doi: 10.1186/s12939-025-02581-5.
- [34] Y. Gao, Y. Feng, and S. G. Jahng, "Implicit bias in digital health: systematic biases in large language models' representation of global public health attitudes and challenges to health equity," *Front. Public Health*, vol. 13, Art. no. 1705082, 2025. doi: 10.3389/fpubh.2025.1705082.
- [35] E. Lopez-Lopez, C. M. Abels, D. Holford, S. M. Herzog, and S. Lewandowsky, "Generative artificial intelligence-mediated confirmation bias in health information seeking," *Ann. N.Y. Acad. Sci.*, vol. 1550, no. 1, pp. 23–36, 2025. doi: 10.1111/nyas.15413.

- [36] G. Currie, J. Currie, S. Anderson, and J. Hewis, "Gender bias in generative artificial intelligence text-to-image depiction of medical students," *Health Educ. J.*, vol. 83, no. 7, pp. 732–746, 2024. doi: 10.1177/00178969241274621.
- [37] M. Achdari et al., "Gender bias in AI's perception of cardiovascular risk," *J. Med. Internet Res.*, vol. 26, p. e54242, 2024. doi: 10.2196/54242.
- [38] G. M. Currie, J. Hewis, E. Hawk, and E. Rohren, "Gender and ethnicity bias of text-to-image generative artificial intelligence in medical imaging, part 1: Preliminary evaluation," *J. Nucl. Med. Technol.*, vol. 52, no. 4, pp. 356–359, 2024. doi: 10.2967/jnmt.124.268332.
- [39] G. M. Currie, J. Hewis, E. Hawk, and E. Rohren, "Gender and ethnicity bias of text-to-image generative artificial intelligence in medical imaging, part 2: Analysis of DALL-E 3," *J. Nucl. Med. Technol.*, vol. 52, no. 4, 2024. doi: 10.2967/jnmt.124.268359.
- [40] T. Templin et al., "Framework for bias evaluation in large language models in healthcare settings," *npj Digit. Med.*, vol. 8, no. 1, 2025. doi: 10.1038/s41746-025-01786-w.
- [41] P. A. Perez-Toro et al., "Exploring biases related to the use of large language models in a multilingual depression corpus," *Sci. Rep.*, vol. 15, no. 1, p. 36197, 2025. doi: 10.1038/s41598-025-19980-x.
- [42] R. Schnepfer et al., "Exploring biases of large language models in the field of mental health: Comparative questionnaire study of the effect of gender and sexual orientation in anorexia nervosa and bulimia nervosa case vignettes," *JMIR Ment. Health*, vol. 12, p. e57986, 2025. doi: 10.2196/57986.
- [43] R. Huang et al., "Evaluation and bias analysis of large language models in generating synthetic electronic health records: comparative study," *J. Med. Internet Res.*, vol. 27, p. e65317, 2025. doi: 10.2196/65317.