Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Assessing bias in the use of large language models in healthcare: a systematic review

Lillian Muyama

muyamalillian@gmail.com

Makerere University https://orcid.org/0000-0002-0409-0176

Andrew Katumba

Makerere University

Joyce Nakatumba-Nabende

Makerere University

Systematic Review

Keywords:

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11592366/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Assessing bias in the use of large language models in healthcare: a systematic review

Lillian Muyama[1*], Andrew Katumba[1], Joyce Nabende[1]

1 Makerere University Centre for Artificial Intelligence (MAK-AI), Makerere University, Kampala, Uganda

* lillian.muyama@mak.ac.ug

Abstract

The rapid advancement of Large Language Models (LLMs) has driven their growing adoption across domains, including healthcare. However, these models risk amplifying existing healthcare disparities. This systematic review synthesizes the existing literature on bias assessment in LLMs applied within the healthcare domain. We searched six databases, namely, Scopus, PubMed, Embase, Web of Science, IEEE Xplore, and the ACM Digital Library, for studies published between January 2020 and May 2025, with the final search conducted on 16 May 2025. Retrieved records were screened according to predefined criteria, resulting in 74 included articles. Methodological quality was assessed using the Joanna Briggs Institute Critical Appraisal Checklist for Analytical Cross-Sectional Studies. Owing to methodological heterogeneity, findings were described narratively. LLMs were applied across various healthcare contexts, most notably in medical question answering. Bias assessment methods ranged from human evaluations to statistical analyses, while only 13.5% of the included studies implemented mitigation strategies, which were primarily prompt-based. Overall, 87.8% of the studies reported evidence of bias, with demographic bias being the most commonly investigated bias type. The included studies were of predominantly high quality with minimal risk of bias. The findings indicate that bias remains prevalent in LLM healthcare outputs, highlighting a critical challenge to their equitable and responsible deployment. Future studies should prioritize robust bias evaluation and the implementation of effective bias mitigation strategies to support the safe integration of LLMs into healthcare.

Author summary

Large Language Models (LLMs) are increasingly being applied to healthcare tasks around the world. However, it is still not clear how well these models perform in different healthcare settings and for different patient groups. In this systematic review, we carefully examined 74 studies that applied LLMs across a wide range of healthcare tasks. We looked at the type of LLMs being used, the kinds of biases being studied, how these biases were evaluated, and any methods to mitigate them. We also reviewed where these studies were conducted, the medical domains they focused on, and whether they reported finding bias. Our findings highlight the need for a well-designed framework to assess bias in LLMs used in healthcare. Such a framework would help to produce more consistent and standardized results and support the development of more effective bias mitigation strategies. We also found that most studies were conducted in high-income countries, showing a lack of research from low-

and middle- income countries. More studies are needed in these regions for a more comprehensive understanding of how LLMs perform in different healthcare systems and populations.

Introduction

Large Language Models (LLMs) [1–3] are powerful Artificial Intelligence (AI) models that have been pre-trained on large amounts of text data. Primarily based on the transformer architecture [4], LLMs are deep neural networks typically consisting of billions or even trillions of parameters. These models can generate human-like text and have been applied to a wide range of tasks, particularly in the field of Natural Language Processing (NLP). Examples of such tasks include question answering (QA) [5, 6], language translation [7, 8], classification [9, 10], information extraction [11, 12], and summarization [13, 14]. Additionally, LLMs can be fine-tuned to adapt to more specialized tasks [15]. While initially designed for text generation, some LLMs have since been extended to handle other modalities, such as images [16, 17] and videos [18].

The issue of bias in AI applications has long been a topic of study [19–21]. Given that LLMs have been pre-trained on massive and sometimes uncurated datasets, there is a significant risk of them generating biased outputs. These biased responses can cause harm and help reinforce problematic stereotypes and the spread of misinformation. Moreover, LLMs have been shown to generate nonsensical content, typically known as hallucinations [22–24]. Various studies have shown that LLMs can perpetuate existing stereotypes and biases, including those related to race [25, 26], gender [27, 28], education level [29], and income [30].

Since the public launch of ChatGPT [31] in November 2022, LLMs have become increasingly pervasive, with applications in domains such as healthcare [32–34], education [35], finance [36, 37], and agriculture [38, 39]. Their rapid integration into everyday life underscores the need to ensure that their outputs are fair, unbiased, and reliable. And yet, many widely used LLMs remain opaque, with little transparency on the data used to train them. This lack of clarity raises concerns about the potential harm and inequity that can be caused by these models.

These concerns are especially pertinent in the healthcare domain, due to its high sensitivity, where any clinical decision made may have serious consequences. However, despite these risks, LLMs can be viewed as tools that patients and members of the public could use for health advice. Various studies have been conducted applying LLMs to healthcare tasks, such as medical QA [6, 40], clinical report generation [41, 42], clinical text summarization [43, 44], and diagnosis and treatment recommendations [10, 45–47]. Nevertheless, recent literature suggests that LLMs are not yet suitable for providing definitive clinical advice, and may in fact compound healthcare inequity among different social groups [48].

In this review, we focus on bias in the application of LLMs within the healthcare sector. While several related studies have been done, their objectives and scope differ from ours. For example, Gallegos *et al.* [49] conducted a general survey on bias and fairness in LLMs, primarily focusing on evaluation and mitigation techniques, while Anoop *et al.* [50] explored affective bias in pre-trained language models. Neither of these studies was a systematic review, nor were they specifically focused on the healthcare sector. We also identified two articles [51, 52] that address bias in LLMs in the healthcare domain. In [51], the authors explored the broader impact of AI on health equity for minority groups, with only limited reference to LLMs. In [52], Omar *et al.* reviewed demographic bias in medical LLMs, focusing on bias types, evaluation techniques, and mitigation

strategies. However, their study was limited to demographic bias, predominantly gender, race/ethnicity, and age, overlooking other forms of bias. Moreover, their analysis primarily examined GPT- and BERT-based models, with minimal consideration of other LLMs, and ultimately included a limited number of studies.

On the other hand, our paper aims to fill these gaps by providing a detailed and systematic overview of bias in LLM outputs within the healthcare domain. We aim to inform and support future researchers in recognizing and mitigating these biases when designing and developing LLM-driven healthcare applications. To structure our writing and data synthesis, we address the following three research questions:

RQ1: What types of bias are being assessed in LLMs within the healthcare domain?

RQ2: What methods are being used to evaluate bias in LLM healthcare applications?

RQ3: What mitigation strategies are being employed to reduce bias in healthcare-related LLM applications?

Materials and methods

This systematic review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [53] method for systematic reviews (see S1 Table) to address the proposed research questions.

Search strategy

A comprehensive literature search was performed across six electronic databases, *i.e.*, Scopus, PubMed, Embase, Web of Science, IEEE Xplore, and ACM Digital Library. The search aimed to identify relevant studies assessing bias in LLMs applied in healthcare contexts. The search terms used were related to LLMs, healthcare, and bias or fairness. Only articles published from 2020 onward were considered, and the last search was conducted on 16 May 2025. This initial search yielded 1,634 articles. A detailed breakdown of the number of records retrieved from each database, along with exact search strings used, is provided in S2 Table.

Eligibility criteria

Studies were considered eligible for inclusion based on the criteria specified in Table 1. Articles were included if they met all the inclusion criteria and excluded if they met any of the exclusion criteria.

Table 1. Inclusion and exclusion criteria for the study selection process.

Inclusion Criteria	Exclusion Criteria
Actively investigates bias in an LLM applied within the healthcare domain	Non-LLM algorithm or non-healthcare domain
Has an element of clinical decision support	Lacks a clinical decision support element (*e.g.,* peer-review of medical abstracts)
Written in English	Non-English Article
Original research study	Non-original research articles
Peer-reviewed article	Non-peer-reviewed article
Specifies bias type(s) and LLM(s) applied	Does not specify bias type(s) or LLM(s) used

Study selection

The study selection process followed the PRISMA framework. All articles retrieved from the database search were independently screened by two authors for inclusion. Any conflicts were discussed with a third author who supervised the selection process. First, duplicate articles were removed. Titles and abstracts were then screened against the eligibility criteria. Articles that passed this stage underwent full-text screening to confirm eligibility.

To broaden our study's coverage, we conducted a backward search by reviewing the references of the prior identified articles, as recommended in [54, 55]. Additionally, we conducted a forward search [55] using Google Scholar and the Semantic Scholar API, to identify articles that had cited the initial included articles. These were also screened based on their titles, abstracts, and full-texts. An additional article that was identified during another study was also included. This entire review process was managed in Google Sheets.

Data extraction and quality assessment

Data extraction was independently conducted by two researchers, and any discrepancies were resolved by discussion. For each included study, the following key information pertinent to the research questions was extracted: first author's last name, year of publication, country of the first author's primary affiliation, motivation of the study, clinical domain investigated, type(s) of bias assessed, nature of input and output data, bias evaluation methods employed, and bias mitigation strategies if reported. All reported results relevant to these outcome domains were extracted from each included study and compiled in a Google Sheet.

Two authors independently assessed the methodological quality of the included studies using the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Analytical Cross-Sectional Studies, which comprises eight items. Any disagreements were resolved through discussion, and when necessary, adjudication by the third author. Studies were subsequently categorized as high, moderate, or low quality based on the number of affirmative ("Yes") responses, where 7-8 indicated high quality, 4-6 moderate quality, and 0-3 low quality.

Data synthesis

Due to the heterogeneity of the included studies, the results of the study were mainly synthesized narratively. Extracted data was standardized where necessary to ensure consistency across studies. Findings were summarized using tables, as well as frequency distributions, including the proportion of studies reporting bias, types of LLMs applied, types of bias identified, and the prevalence of the different bias assessment methods and mitigation strategies.

LLMs and bias

The term "Large Language Model" generally refers to language models with billions of parameters. While there is no exact agreed threshold number of parameters, in this study, we adopt the definition stated in [3], which defines an LLM as a pre-trained language model trained on a large corpus of text to learn generic representations that are transferable across various NLP tasks, typically comprising tens to hundreds of billions of parameters. LLMs can be broadly categorized along

two key dimensions: *ownership* and *scope*.

1. **Ownership**

    - *Proprietary LLMs* are developed and owned by private companies. Their training data, model weights, and development methods remain private, offering limited transparency into their functioning. However, they often achieve better performance than open-source LLMs due to the extensive resources utilized in training and model optimization.
    - *Open-source LLMs*, in contrast, make information such as model weights publicly available. This allows for local deployment and fine-tuning, which is particularly valuable in sensitive domains such as healthcare that have valid concerns about data privacy and security.

2. **Scope**

    - *General-purpose LLMs* are trained on large and diverse data, and are capable of performing a wide variety of tasks across multiple domains.
    - *Domain-specific LLMs* are pre-trained or fine-tuned for application in a specific domain, such as healthcare, finance, and law, among others. In the context of this review, domain-specific LLMs refer to those that have been particularly designed for healthcare applications.

In the context of LLM applications in healthcare, we define bias as any unfair or unjustified difference in the outputs generated by the same LLM under comparable conditions across different patient groups.

Results

The initial search resulted in 1,634 articles. After deduplication and screening, 51 articles met the eligibility criteria. The backward search on these articles resulted in the identification of 5 additional eligible articles after article screening. The forward search on the 56 articles (initial and backward search results) resulted in 17 further articles. We also included an additional article that was identified during another study, resulting in a total of 74 included studies in this review, as presented in the PRISMA flow diagram in Figure 1.

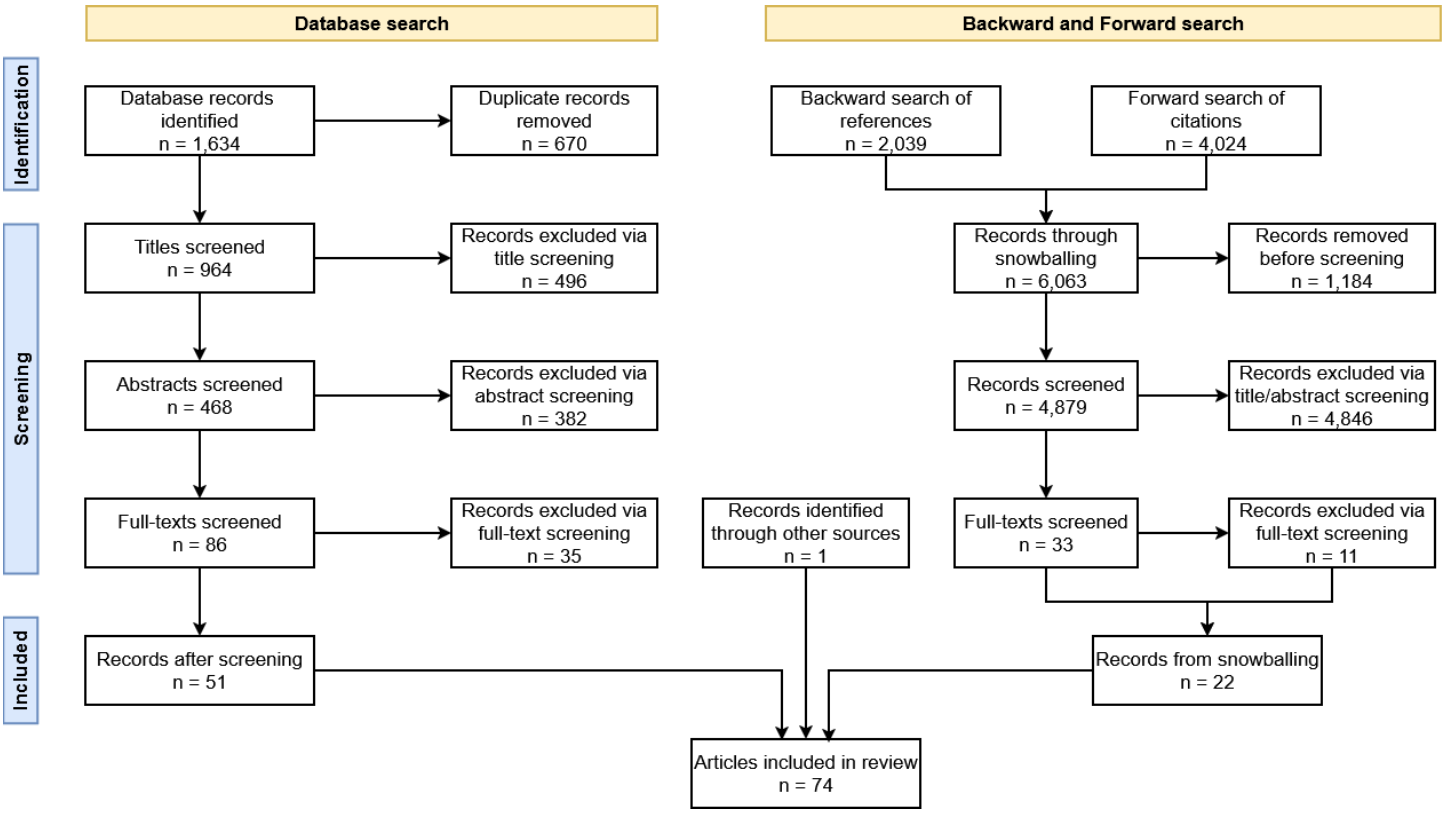


Fig 1. PRISMA flow diagram.

Study quality assessment showed that most of the studies (n=52) were of high quality, with only 3 studies deemed low quality. Overall, the literature demonstrated a low risk of bias. Findings were synthesized descriptively at the study level, across medical domains, LLM types, bias types, assessment methods, and mitigation strategies, among others. A summary of the included studies is presented in Table 2, and the results of the study quality assessment are provided in S3 Table.

Table 2. Summary of included articles. Abbreviations: G/S=Gender/Sex; R/E=Race/Ethnicity; L/C=Location/Culture; E/L=Education background/Literacy level.

Ref.	Country	Task	LLMs	Bias Type	Bias Evaluation	Bias Found	Bias Mitigation
[6]	USA	Medical QA/Chatbot	Med-PaLM	G/S, R/E	Human experts	✗	N/A
[25]	USA	Risk prediction	GPT	R/E	Subgroup performance comparison, Statistical testing	✓	✗
[32]	USA	Medical QA/Chatbot	PaLM, Flan-PaLM, Med-PaLM	G/S	Human experts	✓	✗
[56]	USA	ICD coding	PLM-ICD	G/S	Other	✓	✗
[57]	USA	Diagnosis	GPT, BERT, T5	G/S, R/E	Statistical testing	✓	✗
[58]	USA	Diagnosis, Clinical trial design, Risk prediction, Named entity recognition	LLaMA, Mistral, Alpaca, Asclepius, Clinical Camel, MedAlpaca	G/S, R/E	Subgroup performance comparison	✓	✗
[59]	USA	Treatment recommendations	GPT, Gemini	G/S, R/E	Regression, Statistical testing	✗	N/A

Ref.	First author country	Task	LLMs	Bias Type	Bias Evaluation	Bias Found	Bias Mitigation
[60]	USA	Diagnosis, Risk prediction	GPT, LLaMA, Claude	G/S, R/E, Age	Fairness evaluation metric	✓	✓
[61]	USA	Radiology report simplification	GPT	R/E	Subgroup performance comparison	✓	✗
[62]	USA	Medical QA/Chatbot	LLaVA, Qwen, RadFM, OpenFlamingo, MedVInT	G/S, R/E, Age	Subgroup performance comparison	✓	✗
[63]	USA	Radiology report labelling	GPT	G/S, R/E, Age	Subgroup performance comparison, Statistical testing	✓	✗
[64]	USA	Medical QA/Chatbot	GPT, LLaMA, Mixtral, PaLM	Cognitive biases	Subgroup performance comparison	✓	✓
[65]	USA	Medical QA/Chatbot	Med-PaLM	R/E, G/S, Age, Socioeconomic status, Ability, E/L, Language, L/C, Sexual orientation, Religion, Body composition, Familial status	Human experts	✓	✗
[66]	USA	Medical QA/Chatbot	GPT, LLaMA, Gemini	G/S, R/E	Subgroup performance comparison, Statistical testing	✓	✗
[67]	USA	Medical QA/Chatbot, Clinical Trial Design	GPT, Gemini, Claude	G/S, R/E, Income level, E/L, Sexual orientation, Ability, Occupation, Housing status	Subgroup performance comparison	✓	✓
[68]	USA	Treatment recommendations	GPT, LLaMA, Palmyra-Med	G/S, Language, E/L	Subgroup performance comparison	✓	✗
[69]	Israel	Treatment recommendations	GPT	G/S, Occupation	Subgroup performance comparison, Statistical testing	✗	N/A
[70]	USA	Treatment recommendations	GPT	G/S, R/E	Statistical testing	✓	✗
[71]	Switzerland	Medical QA/Chatbot	GPT, LLaMA	G/S, Sexual orientation	Regression	✓	✗
[72]	USA	Treatment recommendations, Diagnosis	GPT, Gemini, Claude, NewMes-15 (LLaMA)	R/E	Human experts	✓	✗
[73]	Taiwan	Treatment recommendations	GPT, Gemini	Therapeutic	Statistical testing	✓	✗
[74]	USA	Diagnosis	LLaMA, Mistral	G/S, R/E	Subgroup performance comparison	✓	✗
[75]	China	Treatment recommendations	GPT, Gemini, Claude	Language	Statistical testing	✓	✗
[76]	Switzerland	Risk prediction	GPT	G/S	Subgroup performance comparison	✓	✗

Ref.	First author country	Task	LLMs	Bias Type	Bias Evaluation	Bias Found	Bias Mitigation
[77]	India	Clinical trial design	GPT	G/S	Subgroup performance comparison, Statistical testing	✗	N/A
[78]	India	Clinical trial design	GPT	G/S, Age	Subgroup performance comparison	✓	✗
[79]	USA	Synthetic data generation	GPT, Gemini, Bing Chat	G/S, R/E	Other	✓	✗
[80]	USA	Medical QA/Chatbot	GPT	G/S, R/E	Subgroup performance comparison, Statistical testing	✓	✗
[81]	USA	Discharge instructions generation	GPT, LLaMA, Gemini	R/E	Statistical testing	✗	N/A
[82]	USA	Medical QA/Chatbot	GPT	G/S, R/E, Sexual orientation	Subgroup performance comparison	✓	✗
[83]	USA	Diagnosis	GPT	G/S, R/E, Age	Regression	✓	✗
[84]	USA	Risk prediction	BERT	G/S, R/E, Income level, Insurance status	Subgroup performance comparison, Statistical testing	✓	✗
[85]	USA	Diagnosis	GPT	G/S, R/E	Regression	✗	N/A
[86]	UAE	Medical QA/Chatbot	GPT, Bard, Bing Chat	R/E, Age, L/C	Human experts	✓	✗
[87]	USA	Medical QA/Chatbot	GPT, LLaMA, Mixtral, Qwen	Language	Subgroup performance comparison	✓	✓
[88]	USA	Medical QA/Chatbot	GPT, LLaMA	R/E, Language	Statistical testing	✓	✗
[89]	USA	Medical QA/Chatbot	GPT	G/S, R/E, L/C	Statistical testing, Regression	✓	✓
[90]	Spain	Medical QA/Chatbot	GPT, LLaMA, Command R+	Language	Subgroup performance comparison	✓	✗
[91]	Japan	Diagnosis, Triage	GPT	R/E	Subgroup performance comparison	✗	N/A
[92]	USA	Diagnosis	GPT, Gemini	R/E	Regression	✓	✗
[93]	China	Treatment recommendations, Synthetic data generation, Triage	Qwen, Ernie, Baichuan	G/S, R/E, Income level, Insurance status, E/L	Subgroup performance comparison, Statistical testing	✓	✗
[94]	USA	Medical QA/Chatbot, Treatment Recommendations, Triage	GPT, LLaMA, Qwen, Gemma, Phi	G/S, R/E, Sexual orientation, Income level	Subgroup performance comparison, Statistical testing	✓	✗
[95]	Brazil	Treatment recommendations	GPT, LLaMA	G/S, Age, E/L	Human experts	✓	✓
[96]	USA	Treatment recommendations	GPT	G/S, Language, E/L	Subgroup performance comparison	✓	✗
[97]	Canada	Treatment recommendations	GPT	Insurance status	Statistical testing	✓	✗
[98]	Israel	Risk prediction	GPT	G/S, L/C	Statistical testing	✓	✗

Ref.	First author country	Task	LLMs	Bias Type	Bias Evaluation	Bias Found	Bias Mitigation
[99]	India	Diagnosis, Risk prediction	GPT	G/S, Age	Subgroup performance comparison	✓	✗
[100]	France	Synthetic data generation	LLaMA, Mistral, BLOOM, Vigogne	G/S	Other	✓	✗
[101]	China	Synthetic data generation	LLaMA, Qwen, Yi	G/S, R/E	Statistical testing, Fairness evaluation metric	✓	✗
[102]	China	Medical QA/Chatbot, Named entity recognition	GPT, Ziya, CHIMED-GPT, Taiyi, Baichuan, MedicalGPT	Mental illness	Other	✓	✗
[103]	USA	Medical QA/Chatbot	GPT	G/S	Subgroup performance comparison	✗	N/A
[104]	Israel	Medical QA/Chatbot	GPT, Claude, Bard	L/C	Statistical testing, Regression	✓	✗
[105]	Israel	Medical QA/Chatbot	GPT, Claude, Bard	L/C	Other	✓	✗
[106]	USA	Medical QA/Chatbot	GPT, Claude, Bard	R/E	Human experts	✓	✗
[107]	USA	Medical QA/Chatbot	GPT	G/S, R/E, Insurance status	Regression	✓	✗
[108]	USA	Medical QA/Chatbot	GPT	Language	Subgroup performance comparison, Statistical testing, Human experts	✓	✗
[109]	UK	Medical QA/Chatbot	GPT, LLaMA, Gemini, Claude	G/S, R/E	Subgroup performance comparison, Fairness evaluation metric	✓	✓
[110]	USA	Medical QA/Chatbot	GPT, Gemini, Co-Pilot	G/S, R/E	Subgroup performance comparison, Statistical testing	✓	✗
[111]	Germany	Medical QA/Chatbot	GPT	G/S, Age	Statistical testing	✓	✗
[112]	USA	Medical QA/Chatbot	GPT, LLaMA	G/S, Age, L/C	Other	✓	✗
[113]	Austria	Medical QA/Chatbot	GPT	Therapeutic	Human experts	✓	✗
[114]	USA	Medical QA/Chatbot	GPT, Bard	G/S, R/E, Income level	Subgroup performance comparison	✓	✗
[115]	China	Medical QA/Chatbot	GPT, Bard, Ernie	Language	Subgroup performance comparison	✓	✗
[116]	Canada	Medical QA/Chatbot	GPT	Language	Statistical testing, Human experts	✓	✗
[117]	USA	Diagnosis	GPT, LLaMA	G/S, R/E	Fairness evaluation metric	✓	✓
[118]	USA	Diagnosis	GPT, LLaVA	G/S, Age	Statistical testing, Fairness evaluation metric	✓	✗

Ref.	First author country	Task	LLMs	Bias Type	Bias Evaluation	Bias Found	Bias Mitigation
[119]	USA	Diagnosis	GPT	G/S, R/E, Age	Subgroup performance comparison, Statistical testing, Regression	✓	✗
[120]	China	Diagnosis	GPT, Qwen	G/S, Age, Disease severity, Record repetition	Other	✓	✓
[121]	China	Diagnosis	GPT	Language	Subgroup performance comparison, Statistical testing	✓	✗
[122]	Turkey	Triage	GPT, Co-Pilot	G/S, Age, Admission time	Subgroup performance comparison	✗	N/A
[123]	USA	Triage	GPT	G/S, R/E, Age, Sexual orientation, Political affiliation	Statistical testing	✓	✗
[124]	France	Triage	LLaMA, Mistral, Mixtral	G/S	Subgroup performance comparison, Statistical testing	✓	✗
[125]	USA	Bias assessment	LLaMA, Mistral, Qwen, Mamba, Pythia, Tulu, Meditron	G/S, R/E, Language	Other	✓	✗
[126]	UK	Synthetic data generation	DALL-E, Midjourney	Skin tone	Statistical testing	✓	✓

General characteristics

As outlined in our search strategy, due to the relative emergence of LLMs, only studies published from 2020 onward were included. Indeed, no studies were identified in 2020 and 2022, and only one study was identified in 2021. The largest proportion of articles (51.4%, n=38) was published in 2024, underscoring the recent surge of interest in applying LLMs in healthcare. Despite the search being conducted only five months into the year, 2025 accounted for the second largest number of articles (36.5%, n=27), further highlighting the growing prevalence of LLM-related research. Figure 2 shows the distribution of included studies by publication year.

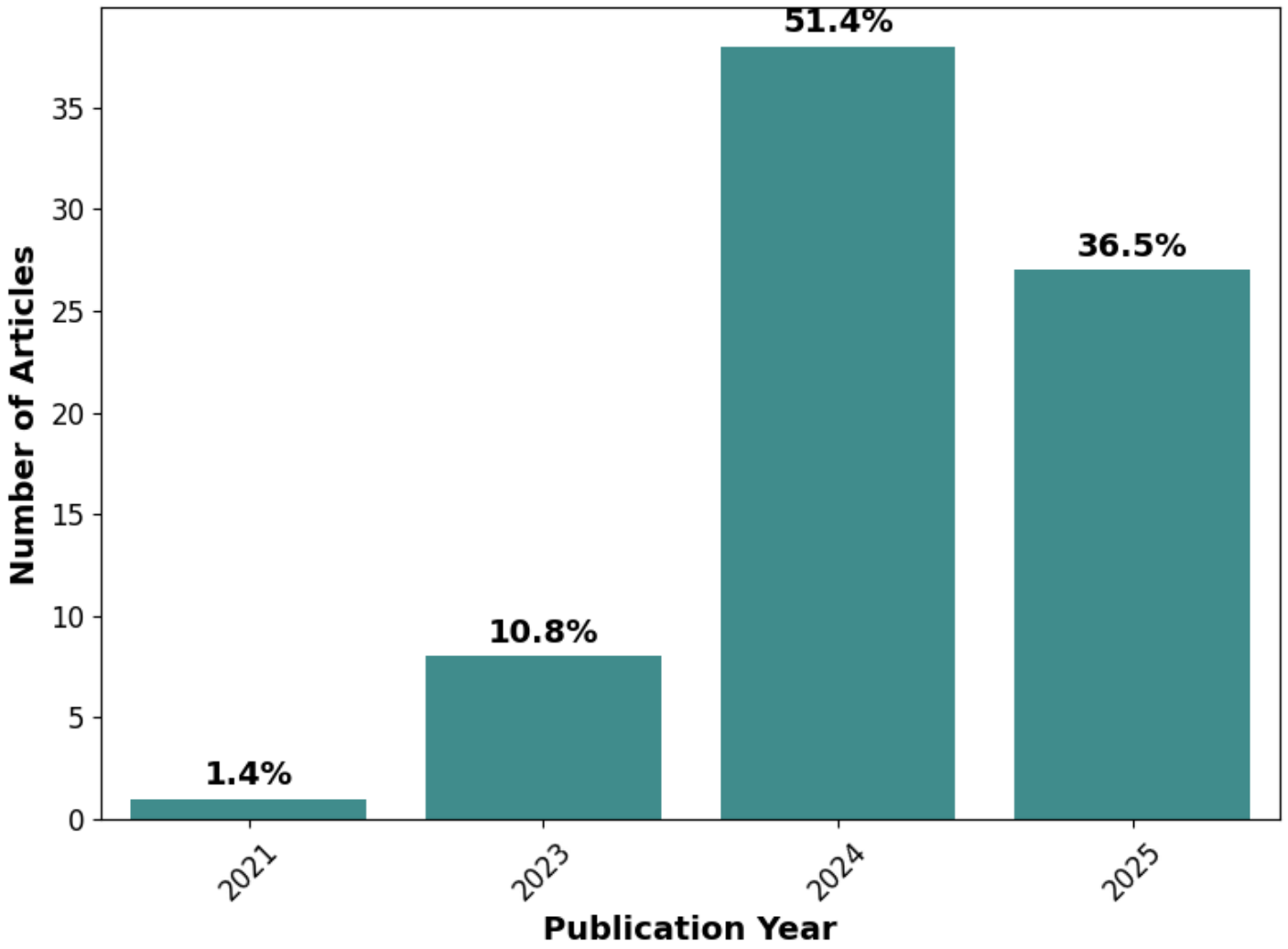


Fig 2. Distribution of included studies by year of publication.

The country of each study's origin was determined based on the first author's primary institutional affiliation. More than half of the included studies (59.5%, n=44) originated from the United States of America (USA), followed by China (9.5%, n=7) and Israel (5.4%, n=4), as depicted in Figure 3.

With respect to datasets, the Medical Information Mart for Intensive Care (MIMIC) dataset series was the most frequently used, appearing in 8 articles. These included MIMIC III [56–58], MIMIC IV [59–61], and MIMIC CXR [62, 63]. The MedQA dataset [127] and its derivatives were also commonly employed, particularly in studies addressing medical QA tasks [6, 32, 60, 64–68].

In terms of medical domains, mental health was the most extensively studied area, with 12 included studies focusing on conditions such as depression [69, 70], eating disorders [71], and psychiatry [72, 73]. Cardiovascular diseases represented the second most common domain, with 5 studies addressing topics, such as arrhythmia and congestive heart failure [74], atrial fibrillation [75], and coronary artery disease [58, 76]. Other domains included oncology [77–79], pain management [59, 80], Human Immunodeficiency Virus (HIV) [81, 82], and nervous system diseases [83–85], among others.

The most common applications of LLMs in healthcare are as chatbots or for medical QA tasks. Overall, LLMs were used as chatbots or for medical QA in 41.9% (n=31) of the included studies. LLMs were also applied to diagnosis in 20.3% (n=15)

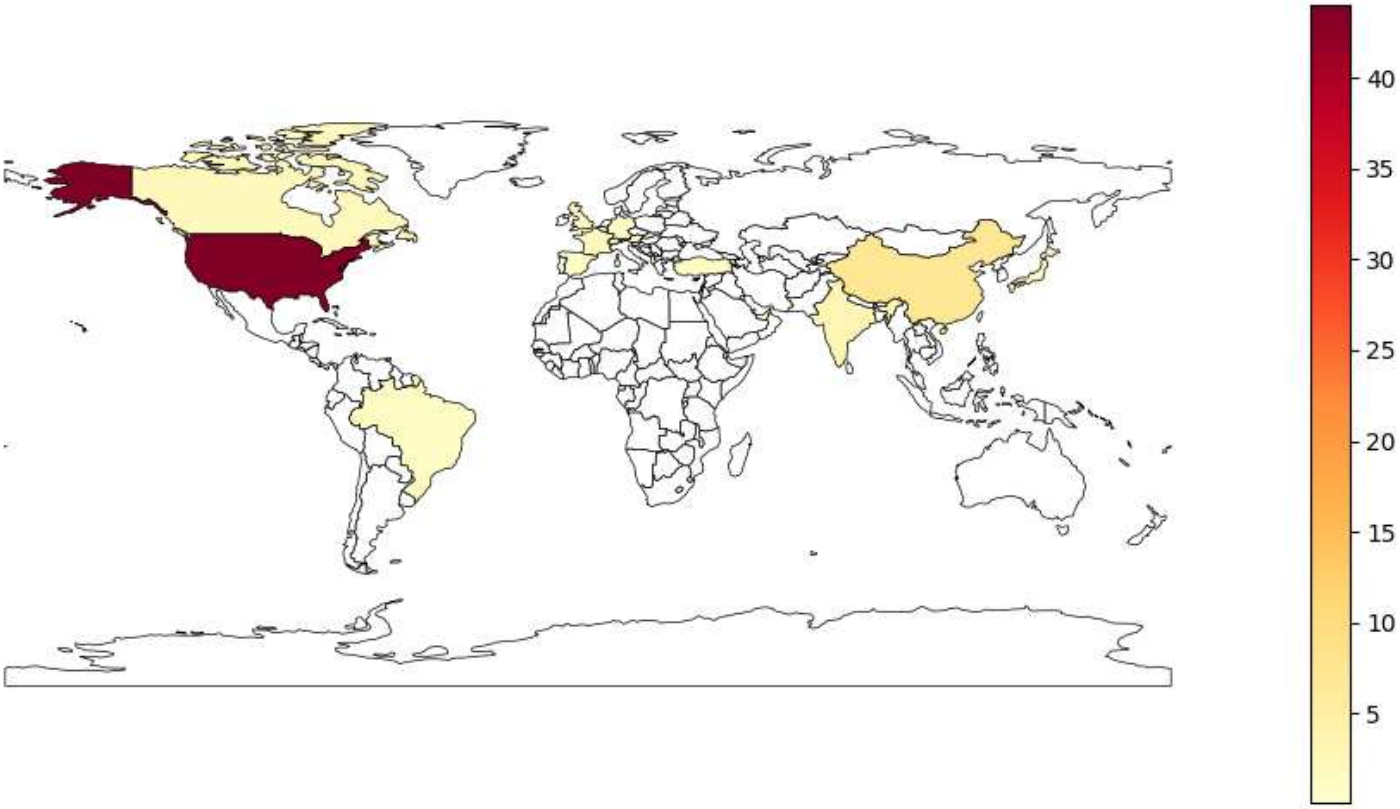


Fig 3. Distribution of included studies by country.

of the included studies. A smaller number of studies applied LLMs to treatment recommendations (16.2%, n=12), while risk prediction accounted for 9.5% (n=7). Other LLM applications in healthcare include clinical trial eligibility assessment [58, 67, 77, 78], medical triage [91, 93, 122–124], synthetic data generation [79, 93, 100, 101], named entity recognition [58, 102], ICD coding [56], and radiology report simplification [61] and labeling [63] (Table 2).

Proprietary LLMs were the most frequently employed, with OpenAI's Generative Pre-trained Transformer (GPT) family [31] dominating usage, appearing in 81.1% (n=60) of the studies. Google's Gemini [128] was used in 14.9% (n=11) of the included studies, while Anthropic's Claude models appeared in 10.4% (n=8) of the included studies. Other proprietary models used include Co-Pilot/Bing Chat [79, 86] and Google Bard [104–106, 114]. Among open-source models, the LLaMA family was the most prominent, applied in 29.9% (n=23) of included studies, while both Mistral and Qwen models appeared in 9.1% (n=7) of the included studies. Domain-specific LLMs were less common, with Med-PaLM and its variants being the most widely applied among those used. Other domain-specific LLMs include CHIMED-GPT and MedicalGPT [102], BioMistral [124], MedAlpaca and Clinical Camel [58], MedFlamingo and MedVInT [62], Meditron [125], NewMes-15 [72], and MentaLLaMA [71].

Overall, 87.8% (n=65) of the included studies reported evidence of bias in at least one LLM on at least one healthcare task. In contrast, 12.2% (n=9) found no evidence of bias in the LLM outputs for their selected tasks.

RQ1: What types of bias are being assessed in LLMs within the healthcare domain?

A summary of the major types of bias investigated in the included studies is presented in Table 3. The most frequently examined biases in the included studies were demographic biases, particularly those related to gender/sex, race/ethnicity, and

age. The most common among these was gender/sex bias, which was assessed in 67.6% (n=50) of the included studies. Although gender and sex represent distinct concepts, with sex referring to biological attributes and gender to social identity, they were grouped together in our review. Race/ethnicity bias was investigated in 52.7% (n=39) of the included studies, while age bias appeared in 21.6% (n=16). For example, Young *et al.* [59] investigated sex and race bias in pain management recommendations generated by GPT-4 and Gemini, while Gao *et al.* [74] assessed similar biases in Mistral and LLaMA's diagnoses. Zhou *et al.* [60] evaluated age, race, and sex biases in several healthcare prediction tasks, including mortality prediction, outcome prediction, and readmission, and Patel *et al.* [83] investigated these same demographic biases in ChatGPT's diagnostic recommendations. Other demographic biases investigated include sexual orientation [65, 67, 71], disability status [65, 67], and religion [65].

Table 3. Types of bias investigated in the included studies. An article may investigate more than one bias type.

Bias type	N (%)	Examples of included studies
Demographic		
Gender/Sex	67.6%	[56, 59, 70, 74, 89, 99, 111, 118]
Race/Ethnicity	52.7%	[25, 57, 60, 81, 83, 88, 94, 117]
Age	21.6%	[60, 86, 95, 111, 112, 118, 119]
Sociocultural		
Language	16.2%	[68, 87, 96, 108, 115, 121, 125]
Geographical location/culture	9.5%	[64, 86, 89, 98, 104, 105, 112]
Socioeconomic		
Education background/Literacy level	8.1%	[67, 68, 93, 95, 96]
Income level	6.8%	[67, 69, 84, 93, 94, 114]
Insurance status	5.4%	[84, 93, 97, 107]
Cognitive	1.4 %	[64]

Biases based on sociocultural factors were the next most common, with language bias appearing in 16.2% (n=12) of included studies. Several included studies investigated the influence of the language used on the recommendations or answers from LLMs. Restrepo *et al.* [87] introduced a multilingual medical QA dataset and found significant differences in LLM responses depending on the input language, while Burn *et al.* [96] observed that responses to non-English queries were shorter and less detailed. Other sociocultural factors include geographical location and culture, assessed in 9.5% (n=7) of included studies. For instance, Zack *et al.* [89] assessed whether the country of origin of a medical educator influenced GPT-4's outputs when generating clinical vignettes. In another study [112], location was limited to states in the US, while Hadar-Shoval *et al.* [104, 105] explored potential cultural biases in LLM clinical recommendations.

Socioeconomic biases were investigated as well, including those related to educational background [93, 95, 96], income level [67, 84, 93, 94, 114], insurance status [84, 93, 97, 107], and occupation [67, 69]. Several studies investigated other forms of bias, such as skin tone bias in the generation of medical images [126], therapeutic bias toward particular treatments [73, 113], and cognitive biases [64].

RQ2: What methods are being used to evaluate bias and fairness in LLM healthcare applications?

The included studies used a variety of strategies to measure and evaluate bias, which are summarized in Table 4. As shown, approaches based on comparing performance across different subgroups were the most common. For example, Restrepo *et*

al. [87] assessed bias by comparing accuracy across languages in LLMs applied to a medical ophthalmological QA task. Schmidgall *et al.* [64] measured changes in accuracy upon the introduction of biased prompts, while Arroyo *et al.* [58] examined mean differences in AUROC (Area under the Receiver Operating Characteristic Curve) and F1 scores for different demographic groups.

Table 4. Bias assessment methods used in the included studies.

Bias assessment method	N (%)	Examples of included studies
Subgroup performance comparison	45.9	[57, 75, 81, 89, 101, 119, 121]
Statistical tests	41.9	[58, 62, 64, 82, 87, 90, 109]
Human experts	13.5	[65, 72, 86, 95, 106, 108]
Regression	12.2	[83, 85, 89, 92, 107, 119]
Fairness evaluation metric	6.8	[60, 101, 109, 117, 118]

Statistical hypothesis testing was another popular approach, with the chi-squared test being the most widely used [57, 81, 119]. Guevara *et al.* [57] applied chi-squared tests and 2-proportion z tests to examine discrepancies in Social Determinants of Health (SDoH) classifications with and without demographic data in the prompt, while Andreadis *et al.* [119] used one-way Analysis of Variance (ANOVA) to explore differences based on various demographic variables, a linear regression model to predict deviation scores using various demographics as independent variables, and chi-squared tests to analyze model responses across different demographics. Other studies that used chi-squared tests include [63, 69, 75, 89, 93, 101, 104, 118], while the chi-squared goodness-of-fit test was applied in [123, 126]. ANOVA was also frequently used, with multivariate two-way ANOVA analysis employed in [98] to evaluate the effect of several demographic variables on model output and one-way ANOVA applied by Dietrich *et al.* [97] to analyze differences in generated model outputs across insurance types. Moreover, [111] and [116] leveraged ANOVA followed by Tukey's tests for comparative analysis. The t-test family appeared in multiple studies as well [77, 80, 94, 119, 121, 124]. Additional statistical hypothesis tests reported include the z-test [25, 57, 66], Fisher's exact test [59, 75, 84, 89], Mann-Whitney test [89, 93], and McNemar's test [70]. For statistical significance, most of the above studies used p-values.

Several studies tasked human experts to rate bias in LLM responses. Aremu *et al.* [86] employed culturally-aware experts in mental health-related disciplines such as psychology and psychiatry to assess evidence of bias in responses generated with demographic details included in prompts. Reis *et al.* [95] employed a linguist to judge potential bias after applying a debiasing prompt, and Pfohl *et al.* [65] gathered input from physicians, health equity experts, and consumers to quantify bias in model responses.

Other studies adopted regression-based approaches to model bias. Patel *et al.* [83] applied logistic regression to examine race/ethnicity bias and Poisson regression to assess differences between LLM responses by age group. In [89], an ordinal logistic regression was used to assess the difference in GPT-4's responses by race/ethnicity and gender. Similarly, Nastasi *et al.* [107] relied on logistic regression to estimate the odds of LLM-generated outcomes based on race, gender, and insurance status.

Finally, a subset of included studies employed fairness evaluation metrics drawn from algorithmic bias evaluation frameworks. Disparate Impact (DI), which compares the proportion of favourable outcomes for an unprivileged group, *e.g.,* a group with a particular demographic, to the proportion of favourable outcomes for a privileged (reference) group was used in [117, 118]. Equal Opportunity difference (EOD), which measures the differences in the True Positive Rate across different

groups, appeared in [60, 117, 118]. Demographic/Statistical Parity Difference, which quantifies differences in outcomes across different groups, was employed in [60, 101], while Average Odds Difference (AOD), which measures the average difference in the True Positive Rate and the False Positive Rate between a privileged and an unprivileged group, was reported in [118]. Equality of Odds (EO) was used in [109] to assess model consistency across different groups for both positive and negative outcomes.

RQ3: What mitigation strategies are being employed to reduce bias in healthcare-related LLM applications?

Once bias was identified, some studies (13.5%, n=10) proposed strategies to mitigate it, as summarized in Table 5.

Table 5. Categories of bias mitigation strategies employed in the included studies.

Mitigation category	Brief description	N (%)	Included studies
Prompt-based	Modifying the input (prompt) to the LLM in such a way as to elicit a fairer or less biased response.	9.5	[60,64,87,89,95,109, 117, 120, 126]
Training-based	Adjusting model parameters during training or post-training.	2.7	[60, 67, 109]
Knowledge augmentation	Providing additional knowledge from external sources, such as web pages and external databases.	1.4	[60, 87]

The most frequently employed approaches were prompt-based methods. For example, Schmidgall *et al.* [64] employed three techniques, *i.e.* bias education, where a warning about potential biases was incorporated into the prompt, one-shot bias demonstration, where a sample question with a bias-inducing prompt and a sample biased answer were presented to the LLM, and few-shot bias demonstration, where both a negative and a positive sample were included in the prompt. Other prompt-based strategies included excluding demographic details (specifically age and gender) from prompts [120], or explicitly instructing models to avoid bias and to consider disease prevalence across demographic groups when making diagnostic recommendations [89]. Benkirane *et al.* [109] implemented two prompt engineering techniques, instruction following (IF) and Chain-of-Thought (CoT) to reduce bias.

A second category of approaches focused on training-based methods, including model training or fine-tuning, which involve adjusting model parameters. Ji *et al.* [67] proposed EquityGuard, a framework leveraging contrastive learning during training to separate socio-demographic factors from task-relevant features, thereby preventing them from influencing model outcomes. In addition to prompt engineering for bias mitigation, Benkirane *et al.* [109] also used Counter-factual Patient Variations (CPVs) to account for patient demographics that may be underrepresented in the datasets used to train the models.

Meanwhile, Restrepo *et al.* [87] developed a framework that integrates multiple specialized agents responsible for prompt translation into English, Retrieval Augmented Generation (RAG), web search, and prompt rewriting, to collaboratively process queries and reduce bias in the final outputs.

Discussion

The majority of studies included in this review (87.8%) reported evidence of bias in LLM-generated responses according to various criteria. However, it is worth asking if the LLM is truly exhibiting bias or providing the most appropriate response based on the given context. For instance, when an LLM offers different recommendations based on a user's insurance status or income level, this may not necessarily constitute discrimination. Rather, it could represent a more achievable solution. In such a case, what appears as bias may actually be contextually and ethically appropriate. Therefore, when evaluating bias in LLM responses, we believe that researchers should distinguish between unfairness or unjust bias and context-aware reasoning. This distinction requires that the framework against which the bias being reported is evaluated is well defined. Additionally, encouraging LLMs to include their reasoning within the given response, for example, using the Chain-of-Thought approach [129], may further aid in this goal.

Some included studies reported bias based on the difference in the performance metric used, such as accuracy. However, it is worth considering that such differences may not necessarily stem from bias, but rather statistical disparities in data. For example, if one demographic group is overrepresented in a dataset, observed differences in LLM response accuracy across demographic groups may reflect data imbalance rather than bias. We, therefore, recommend that in such cases, established fairness metrics such as disparity impact or parity difference could be employed in addition for a more rigorous evaluation.

Furthermore, prior research has shown that users tend to overly trust LLM outputs [130, 131], often relying on them more than appropriate. However, given that a large proportion of studies included in this review report bias in LLM responses, caution should be taken when applying LLMs for healthcare purposes. Models should be designed to communicate their degree of uncertainty in their output, and to remind users to verify information or consult healthcare professionals as necessary. It is also important that the key principles of responsible AI, that is, Fairness, Transparency, Reliability and Safety, Privacy and Security, and Accountability, are systematically integrated into model development and deployment. A human-in-the-loop framework is likewise essential to ensure human oversight and accountability in decision-making processes, as LLMs cannot evaluate the quality of their responses themselves.

Additionally, several studies used publicly available datasets as reported in the Results section. However, a limitation arises when assessing proprietary LLMs, whose training datasets are undisclosed. If such a model has already been trained on this public dataset that is being used in the assessment, this could have an impact on the results reported. Transparency regarding LLM training data sources is therefore critical for meaningful bias assessment.

Research gaps

Lack of studies from LMICs. As illustrated in Figure 3, the majority of the included studies originate from High-Income Countries (HICs). This imbalance raises concerns regarding the generalizability of current findings, especially given the fact that LLMs are largely trained on Western-centric data. Consequently, their performance in low- and middle-income countries, particularly with regard to bias and fairness, remains insufficiently understood. Future research should prioritize evaluating LLM behaviour across diverse cultural, linguistic, and geographical contexts to better understand and address potential biases and disparities in model outputs.

Narrow coverage of geographic and sociocultural bias. While plenty of studies have investigated demographic bias in LLM outputs, the scope of bias assessment remains limited. In particular, there is comparatively little research addressing biases related to geographic location and broader sociocultural contexts. Within the included studies in this review, only one explicitly examined bias based on the intended country of use. This underscores the need for further research into geographic and sociocultural biases beyond language. Even with respect to language, only a limited number of languages have been explored. Moreover, although gender, age, and race are commonly studied, other demographic attributes remain underexplored and future research should grant them more attention.

Insufficient research on domain-specific LLMs. Additionally, this review identified a relative lack of studies on bias in domain-specific LLMs. Further research is warranted to assess whether domain-specific LLMs exhibit similar or different bias patterns compared to general-purpose LLMs. For example, one of the healthcare-specific LLMs in this review, MedPaLM, may offer significant benefits over general-purpose LLMs like GPT for clinical tasks. However, there needs to be careful independent bias evaluation as this may have larger implications on health equity and other aspects of patient care.

Limited development of bias mitigation strategies. Another notable research gap concerns the development of bias mitigation strategies. Only a small portion of the included studies in this review (13.5%) implemented a mitigation approach. Given the increasing integration of LLMs in various domains, including critical ones like healthcare, the development and comprehensive evaluation of bias mitigation strategies should be a priority for future research. Among the studies that implemented bias mitigation strategies, the approaches were mainly either prompt-based or training-based. In studies that employed both techniques [60, 109], they were applied independently. Future work could explore combining these two approaches and evaluating the effectiveness of that combination compared to using each method in isolation.

Development of a standardized bias assessment framework. Finally, the included studies in this review employed a wide range of methods for bias assessment in LLMs. This heterogeneity of evaluation methods not only makes it difficult to do comparisons across studies but also limits the ability to draw broader conclusions about bias prevalence in LLMs. We therefore advocate for the development of a standardized and systematic framework for bias assessment in LLMs, which can be leveraged in future studies. We believe that such a framework will facilitate more consistent comparisons across studies as well as contribute to a clearer understanding of the extent of bias in these applications. This, in turn, will support the development of more effective bias mitigation strategies and ultimately aid the responsible integration of LLMs into the healthcare domain.

Limitations of the study

While the objective of this review article is to provide a comprehensive overview of bias assessment in the application of LLMs in healthcare, we acknowledge that it has some limitations.

Firstly, GPT was included in the query string used to retrieve articles from the identified databases (see S2 Table). We realize this may have biased the results towards articles using GPT-based LLMs, which accounted for over 80% of the included studies. However, we hypothesize that articles using other LLMs were more likely to mention the term "large language model" in their title or abstract. That notwithstanding, we believe the included studies adequately answer the research questions of the study.

Secondly, article selection by title, abstract, and full-text was based on the authors' understanding and perspectives. This influenced the final selection of papers, potentially introducing some bias in article selection.

Conclusion

This review provides a comprehensive examination of bias assessment in LLMs applied within the healthcare domain, summarizing the current state of research, identifying potential areas of improvement, and proposing directions for future work. Our analysis indicates that most LLMs, when applied to a healthcare task, exhibit some form of bias. Although existing studies have primarily focused on demographic biases, such as those related to gender/sex, race/ethnicity, and age, there is comparatively limited research on geographic and cultural bias. The absence of studies conducted in LMICs further highlights the need to evaluate LLM performance in more diverse and underrepresented contexts.

To aid comparison and generalizability of findings across studies, we advocate for the development of a standardized framework for bias assessment in healthcare-focused LLMs. At the same time, LLMs hold promise for a range of healthcare applications, from medical education to clinical decision-making. However, given the documented presence of bias in their outputs, it is essential that these technologies are developed and deployed responsibly, promoting fairness, accountability, transparency, security, robustness, and human-centered design.

Supporting information

S1 Table. PRISMA checklist. (PDF)

S2 Table. Full search strings and number of retrieved articles for each database. (PDF)

S3 Table. Individual study quality assessment. (PDF)

CRediT authorship contribution statement

Lillian Muyama: Conceptualization, Methodology, Investigation, Validation, Writing - original draft, Writing - Review & Editing.

Andrew Katumba: Conceptualization, Investigation, Validation, Resources, Writing - Review & Editing, Supervision, Project Administration, Funding acquisition.

Joyce Nabende: Conceptualization, Investigation, Validation, Resources, Writing - Review & Editing, Supervision, Project Administration, Funding acquisition.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

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

2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in Neural Information Processing Systems. 2020;33:1877–1901.

3. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology. 2023;.

4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30.

5. Lehmann J, Meloni A, Motta E, Osborne F, Recupero DR, Salatino AA, et al. Large language models for scientific question answering: An extensive analysis of the SciQA benchmark. In: European Semantic Web Conference. Springer; 2024. p. 199–217.

6. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nature Medicine. 2025;31(3):943–950.

7. Ji B, Duan X, Zhang Y, Wu K, Zhang M. Zero-shot prompting for LLM-based machine translation using in-domain target sentences. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2024;.

8. Gong J, Foo LG, He Y, Rahmani H, Liu J. LLMs are good sign language translators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024. p. 18362–18372.

9. Sun X, Li X, Li J, Wu F, Guo S, Zhang T, et al. Text Classification via Large Language Models. In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023. p. 8990–9005. Available from: `https://aclanthology.org/2023.findings-emnlp.603/`.

10. Castagnari E, Muyama L, Coulet A. Prompting large language models for supporting the differential diagnosis of anemia. In: 2024 2nd International Cconference on Foundation and Large Language Models (FLLM). IEEE; 2024. p. 253–261.

11. Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, et al. Structured information extraction from scientific text with large language models. Nature communications. 2024;15(1):1418.

12. Xu D, Chen W, Peng W, Zhang C, Xu T, Zhao X, et al. Large language models for generative information extraction: A survey. Frontiers of Computer Science. 2024;18(6):186357.

13. Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto TB. Benchmarking large language models for news summarization. Transactions of the Association for Computational Linguistics. 2024;12:39–57.

14. Zhang H, Yu PS, Zhang J. A systematic survey of text summarization: From statistical methods to large language models. ACM Computing Surveys. 2025;57(11):1–41.

15. Yang H, Zhang Y, Xu J, Lu H, Heng PA, Lam W. Unveiling the Generalization Power of Fine-Tuned Large Language Models. In: Duh K, Gomez H, Bethard S, editors. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico: Association for Computational Linguistics; 2024. p. 884–899. Available from: https://aclanthology.org/2024.naacl-long.51/.

16. Koh JY, Fried D, Salakhutdinov RR. Generating images with multimodal language models. Advances in Neural Information Processing Systems. 2023;36:21487–21506.

17. Brade S, Wang B, Sousa M, Oore S, Grossman T. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology; 2023. p. 1–14.

18. Kondratyuk D, Yu L, Gu X, Lezama J, Huang J, Schindler G, et al. VideoPoet: a large language model for zero-shot video generation. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24. JMLR.org; 2024.

19. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine. 2018;178(11):1544–1547.

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

21. Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal ME, et al. Bias in data-driven artificial intelligence systems-An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2020;10(3):e1356.

22. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024;630(8017):625–630.

23. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys. 2023;55(12). doi:10.1145/3571730.

24. Raunak V, Menezes A, Junczys-Dowmunt M. The Curious Case of Hallucinations in Neural Machine Translation. In: Toutanova K, Rumshisky A, Zettlemoyer L, Hakkani-Tur D, Beltagy I, Bethard S, et al., editors. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics; 2021. p. 1172–1183. Available from: https://aclanthology.org/2021.naacl-main.92/.

25. Yang Y, Liu X, Jin Q, Huang F, Lu Z. Unmasking and quantifying racial bias of large language models in medical report generation. Communications medicine. 2024;4(1):176.

26. Bai X, Wang A, Sucholutsky I, Griffiths TL. Explicitly unbiased large language models still form biased associations. Proceedings of the National Academy of Sciences. 2025;122(8):e2416228122.

27. Fulgu RA, Capraro V. Surprising gender biases in GPT. Computers in Human Behavior Reports. 2024;16:100533.

28. Ho JQ, Hartanto A, Koh A, Majeed NM. Gender biases within Artificial Intelligence and ChatGPT: Evidence, sources of biases and solutions. Computers in Human Behavior: Artificial Humans. 2025; p. 100145.

29. Qu Y, Wang J. Performance and biases of large language models in public opinion simulation. Humanities and Social Sciences Communications. 2024;11(1):1–13.

30. Weissburg I, Anand S, Levy S, Jeong H. LLMs are Biased Teachers: Evaluating LLM Bias in Personalized Education. In: Chiruzzo L, Ritter A, Wang L, editors. Findings of the Association for Computational Linguistics: NAACL 2025. Albuquerque, New Mexico: Association for Computational Linguistics; 2025. p. 5650–5698. Available from: `https://aclanthology.org/2025.findings-naacl.314/`.

31. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv preprint arXiv:230308774. 2023;.

32. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–180.

33. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature Medicine. 2023;29(8):1930–1940.

34. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: Development, applications, and challenges. Health Care Science. 2023;2(4):255–263.

35. Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences. 2023;103:102274.

36. Wu S, Irsoy O, Lu S, Dabravolski V, Dredze M, Gehrmann S, et al. Bloomberggpt: A large language model for finance. arXiv preprint arXiv:230317564. 2023;.

37. Li Y, Wang S, Ding H, Chen H. Large language models in finance: A survey. In: Proceedings of the fourth ACM international conference on AI in finance; 2023. p. 374–382.

38. Tzachor A, Devare M, Richards C, Pypers P, Ghosh A, Koo J, et al. Large language models and agricultural extension services. Nature Food. 2023;4(11):941–948.

39. Rezayi S, Liu Z, Wu Z, Dhakal C, Ge B, Dai H, et al. Exploring new frontiers in agricultural NLP: Investigating the potential of large language models for food applications. IEEE Transactions on Big Data. 2024;.

40. Restrepo D, Nakayama LF, Dychiao RG, Wu C, McCoy LG, Artiaga JC, et al.. Seeing Beyond Borders: Evaluating LLMs in Multilingual Ophthalmological Question Answering; 2024.

41. Nakaura T, Yoshida N, Kobayashi N, Shiraishi K, Nagayama Y, Uetani H, et al. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. Japanese Journal of Radiology. 2024;42(2):190–200.

42. Voinea SV, Mămuleanu M, Teică RV, Florescu LM, Selișteanu D, Gheonea IA. GPT-driven radiology report generation with fine-tuned llama 3. Bioengineering. 2024;11(10):1043.

43. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nature Medicine. 2024;30(4):1134–1142.

44. Bednarczyk L, Reichenpfader D, Gaudet-Blavignac C, Ette AK, Zaghir J, Zheng Y, et al. Scientific evidence for clinical text summarization using large language models: scoping review. Journal of Medical Internet Research. 2025;27:e68998.

45. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. The Lancet Digital Health. 2024;6(8):e555–e561. doi:10.1016/S2589-7500(24)00097-9.

46. Su H, Sun Y, Li R, Zhang A, Yang Y, Xiao F, et al. Large Language Models in Medical Diagnostics: Scoping Review With Bibliometric Analysis. Journal of Medical Internet Research. 2025;27:e72062.

47. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. Health Promotion Perspectives. 2023;13(3):183.

48. Liu F, Li Z, Zhou H, Yin Q, Yang J, Tang X, et al. Large Language Models Are Poor Clinical Decision-Makers: A Comprehensive Benchmark. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 13696–13710. Available from: `https://aclanthology.org/2024.emnlp-main.759/`.

49. Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, et al. Bias and Fairness in Large Language Models: A Survey. Computational Linguistics. 2024;50(3):1097–1179. doi:10.1162/coli_a_00524.

50. Anoop K , Manjary P Gangan, Deepak P , Lajish V L . Towards an Enhanced Understanding of Bias in Pre-trained Neural Language Models: A Survey with Special Emphasis on Affective Bias. In: Mathew J, Santhosh Kumar G, P D, Jose JM, editors. Responsible Data Science. Singapore: Springer Nature; 2022. p. 13–45.

51. Iloanusi NJ, Chun SA. AI impact on health equity for marginalized, racial, and ethnic minorities. In: Proceedings of the 25th annual international conference on digital government research; 2024. p. 841–848.

52. Omar M, Sorin V, Agbareia R, Apakama DU, Soroush A, Sakhuja A, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. International Journal for Equity in Health. 2025;24(1):57.

53. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj. 2021;372.

54. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ. 2005;331(7524):1064–1065. doi:10.1136/bmj.38636.593461.68.

55. Webster J, Watson RT. Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly. 2002; p. xiii–xxiii.

56. Jaganathan GS, Kahanda I, Kanewala U. Metamorphic Testing for robustness and fairness evaluation of LLM-based automated ICD coding applications. Smart Health. 2025; p. 100564.

57. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, et al. Large language models to identify social determinants of health in electronic health records. NPJ digital medicine. 2024;7(1):6.

58. Ceballos-Arroyo AM, Munnangi M, Sun J, Zhang K, Mcinerney J, Wallace BC, et al. Open (clinical) LLMs are sensitive to instruction phrasings. In: Proceedings of the 23rd Workshop on Biomedical Natural Language Processing; 2024. p. 50–71.

59. Young CC, Enichen E, Rao A, Succi MD. Racial, ethnic, and sex bias in large language model opioid recommendations for pain management. Pain. 2022; p. 10–1097.

60. Zhou Y, Di Eugenio B, Cheng L. Unveiling Performance Challenges of Large Language Models in Low-Resource Healthcare: A Demographic Fairness Perspective. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Eugenio BD, Schockaert S, editors. Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: Association for Computational Linguistics; 2025. p. 7266–7278. Available from: `https://aclanthology.org/2025.coling-main.485/`.

61. Amin KS, Forman HP, Davis MA. Even with ChatGPT, race matters. Clinical Imaging. 2024;109:110113.

62. Xia P, Chen Z, Tian J, Gong Y, Hou R, Xu Y, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. Advances in Neural Information Processing Systems. 2024;37:140334–140365.

63. Santomartino SM, Zech JR, Hall K, Jeudy J, Parekh V, Yi PH. Evaluating the performance and bias of natural language processing tools in labeling chest radiograph reports. Radiology. 2024;313(1):e232746.

64. Schmidgall S, Harris C, Essien I, Olshvang D, Rahman T, Kim JW, et al. Evaluation and mitigation of cognitive biases in medical language models. npj Digital Medicine. 2024;7(1):295.

65. Pfohl SR, Cole-Lewis H, Sayres R, Neal D, Asiedu M, Dieng A, et al. A toolbox for surfacing health equity harms and biases in large language models. Nature Medicine. 2024;30(12):3590–3600.

66. Rawat R, McBride H, Nirmal D, Ghosh R, Moon J, Alamuri D, et al. DiversityMedQA: A benchmark for assessing demographic biases in medical diagnosis using large language models. In: Proceedings of the Third Workshop on NLP for Positive Impact; 2024. p. 334–348.

67. Ji Y, Ma W, Sivarajkumar S, Zhang H, Sadhu EM, Li Z, et al. Mitigating the risk of health inequity exacerbated by large language models. npj Digital Medicine. 2025;8(1):246.

68. Gourabathina A, Gerych W, Pan E, Ghassemi M. The medium is the message: How non-clinical information shapes clinical decisions in LLMs. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency; 2025. p. 1805–1828.

69. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Family Medicine and Community Health. 2023;11(4):e002391.

70. Perlis RH, Goldberg JF, Ostacher MJ, Schneck CD. Clinical decision support for bipolar depression using large language models. Neuropsychopharmacology. 2024;49(9):1412–1416.

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

72. Bouguettaya A, Stuart EM, Aboujaoude E. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. npj Digital Medicine. 2025;8(1):332.

73. Chang Y, Liu YC, Huang SS, Hsu WY. Assessing bias in AI-driven psychiatric recommendations: A comparative cross-sectional study of chatbot-classified and CANMAT 2023 guideline for adjunctive therapy in difficult-to-treat depression. Psychiatry Research. 2025;348:116501.

74. Gao Y, Myers S, Chen S, Dligach D, Miller T, Bitterman DS, et al. Uncertainty estimation in diagnosis generation from large language models: next-word probability is not pre-test probability. JAMIA open. 2025;8(1):ooae154.

75. Li Z, Yan C, Cao Y, Gong A, Li F, Zeng R. Evaluating performance of large language models for atrial fibrillation management using different prompting strategies and languages. Scientific Reports. 2025;15(1):19028.

76. Achtari M, Salihu A, Muller O, Abbé E, Clair C, Schwarz J, et al. Gender Bias in AI's Perception of Cardiovascular Risk. Journal of Medical Internet Research. 2024;26:e54242.

77. Devi A, Uttrani S, Singla A, Jha S, Dasgupta N, Natarajan S, et al. Automating clinical trial eligibility screening: Quantitative analysis of GPT models versus human expertise. In: Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments; 2024. p. 626–632.

78. Devi A, Uttrani S, Singla A, Jha S, Dasgupta N, Natarajan S, et al. Quantitative analysis of GPT-4 model: Optimizing patient eligibility classification for clinical trials and reducing expert judgment dependency. In: Proceedings of the 2024 8th International Conference on Medical and Health Informatics; 2024. p. 230–237.

79. Agrawal A. Fairness in AI-Driven Oncology: Investigating Racial and Gender Biases in Large Language Models. Cureus. 2024;16(9).

80. Logé C, Ross E, Dadey D, Jain S, Saporta A, Ng A, et al. Q-Pain: A Question Answering Dataset to Measure Social Bias in Pain Management. Advances in Neural Information Processing Systems. 2021;34.

81. Hanna JJ, Wakene AD, Johnson AO, Lehmann CU, Medford RJ. 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. 2025;27:e57257.

82. Criss S, Nguyen TT, Gonzales SM, Lin B, Kim M, Makres K, et al. "HIV Stigma Exists"-Exploring ChatGPT's HIV Advice by Race and Ethnicity, Sexual Orientation, and Gender Identity. Journal of Racial and Ethnic Health Disparities. 2024; p. 1–14.

83. Patel MA, Villalobos F, Shan K, Tardo LM, Horton LA, Sguigna PV, et al. Generative artificial intelligence versus clinicians: Who diagnoses multiple sclerosis faster and with greater accuracy? Multiple sclerosis and related disorders. 2024;90:105791.

84. Xie K, Ojemann WK, Gallagher RS, Shinohara RT, Lucas A, Hill CE, et al. Disparities in seizure outcomes revealed by large language models. Journal of the American Medical Informatics Association. 2024;31(6):1348–1355.

85. Shan K, Patel MA, McCreary M, Punnen TG, Villalobos F, Tardo LM, et al. Faster and better than a physician?: Assessing diagnostic proficiency of ChatGPT in misdiagnosed individuals with neuromyelitis optica spectrum disorder. Journal of the Neurological Sciences. 2025;468:123360.

86. Aremu T, Akinwehinmi O, Nwagu C, Ahmed SI, Orji R, Amo PAD, et al. On the reliability of Large Language Models to misinformed and demographically informed prompts. AI Magazine. 2025;46(1):e12208.

87. Restrepo D, Wu C, Tang Z, Shuai Z, Phan TNM, Ding JE, et al. Multi-OphthaLingua: A Multilingual Benchmark for Assessing and Debiasing LLM Ophthalmological QA in LMICs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39; 2025. p. 28321–28330.

88. Lee Y, Chang CH, Yang CC. Enhancing Patient-Physician Communication: Simulating African American Vernacular English in Medical Diagnostics with Large Language Models. Journal of healthcare informatics research. 2025;9(2):119–153.

89. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, 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. 2024;6(1):e12–e22.

90. Schlicht IB, Zhao Z, Sayin B, Flek L, Rosso P. Do LLMs provide consistent answers to health-related questions across languages? In: European Conference on Information Retrieval. Springer; 2025. p. 314–322.

91. Ito N, Kadomatsu S, Fujisawa M, Fukaguchi K, Ishizawa R, Kanda N, et al. The accuracy and potential racial and ethnic biases of GPT-4 in the diagnosis and triage of health conditions: evaluation study. JMIR Medical Education. 2023;9:e47532.

92. Deb B, Rodman A. Racial differences in pain assessment and false beliefs about race in AI models. JAMA Network Open. 2024;7(10):e2437977–e2437977.

93. Liu C, Zheng J, Liu Y, Wang X, Zhang Y, Fu Q, et al. Potential to perpetuate social biases in health care by Chinese large language models: a model evaluation study. International Journal for Equity in Health. 2025;24(1):206.

94. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. Nature Medicine. 2025; p. 1–9.

95. Reis ZSN, Pagano AS, de Oliveira IJR, dos Santos Dias C, Lage EM, Mineiro EF, et al. Evaluating large language model-supported instructions for medication use: First steps toward a comprehensive model. Mayo Clinic Proceedings: Digital Health. 2024;2(4):632–644.

96. Burns C, Bakaj A, Berishaj A, Hristidis V, Deak P, Equils O. Use of generative AI for improving health literacy in reproductive health: case study. JMIR Formative Research. 2024;8(1):e59434.

97. Dietrich N, Stubbert B. Abstract No. 96 Does Patient Insurance Influence ChatGPT's Recommendations? A Closer Look at AI Bias in Interventional Radiology. Journal of Vascular and Interventional Radiology. 2025;36(3):S44–S45.

98. Levkovich I, Shinan-Altman S, Elyoseph Z. Can large language models be sensitive to culture suicide risk assessment? Journal of Cultural Cognitive Science. 2024;8(3):275–287.

99. Soun RS, Nair A. ChatGPT for Mental Health Applications: A study on biases. In: Proceedings of the Third International Conference on AI-ML Systems; 2023. p. 1–5.

100. Ducel F, Hiebel N, Ferret O, Fort K, Névéol A. "Women do not have heart attacks!" Gender Biases in Automatically Generated Clinical Cases in French. In: Findings of the Association for Computational Linguistics: NAACL 2025; 2025. p. 7145–7159.

101. Huang R, Wu H, Yuan Y, Xu Y, Qian H, Zhang C, et al. Evaluation and Bias Analysis of Large Language Models in Generating Synthetic Electronic Health Records: Comparative Study. Journal of Medical Internet Research. 2025;27:e65317.

102. Tian Y, Gan R, Song Y, Zhang J, Zhang Y. ChiMed-GPT: A Chinese Medical Large Language Model with Full Training Regime and Better Alignment to Human Preferences. In: Ku LW, Martins A, Srikumar V, editors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics; 2024. p. 7156–7173. Available from: https://aclanthology.org/2024.acl-long.386/.

103. Deva R, Ramani D, Divate T, Jalota S, Ismail A. "Kya family planning after marriage hoti hai?": Integrating Cultural Sensitivity in an LLM Chatbot for Reproductive Health. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; 2025. p. 1–23.

104. Hadar-Shoval D, Asraf K, Shinan-Altman S, Elyoseph Z, Levkovich I. Embedded values-like shape ethical reasoning of large language models on primary care ethical dilemmas. Heliyon. 2024;10(18).

105. Hadar-Shoval D, Asraf K, Mizrachi Y, Haber Y, Elyoseph Z. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic values. JMIR Mental Health. 2024;11:e55988.

106. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. NPJ Digital Medicine. 2023;6(1):195.

107. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. Scientific Reports. 2023;13(1):17885.

108. Jin Y, Chandra M, Verma G, Hu Y, De Choudhury M, Kumar S. Better to ask in English: Cross-lingual evaluation of large language models for healthcare queries. In: Proceedings of the ACM Web Conference 2024; 2024. p. 2627–2638.

109. Benkirane K, Kay J, Perez-Ortiz M. How Can We Diagnose and Treat Bias in Large Language Models for Clinical Decision-Making? In: Chiruzzo L, Ritter A, Wang L, editors. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Albuquerque, New Mexico: Association for Computational Linguistics; 2025. p. 2263–2288. Available from: `https://aclanthology.org/2025.naacl-long.114/`.

110. Lee GG, Goodman D, Chang TCP. Impact of Demographic Modifiers on Readability of Myopia Education Materials Generated by Large Language Models. Clinical Ophthalmology. 2024; p. 3591–3604.

111. Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H. "Doctor ChatGPT, can you help me?" The patient's perspective: cross-sectional study. Journal of Medical Internet Research. 2024;26:e58831.

112. Levy S, Karver T, Adler W, Kaufman M, Dredze M. Evaluating biases in context-dependent sexual and reproductive health questions. In: Findings of the Association for Computational Linguistics: EMNLP 2024; 2024. p. 5801–5812.

113. Raile P. The usefulness of ChatGPT for psychotherapists and patients. Humanities and Social Sciences Communications. 2024;11(1):1–8.

114. Kim J, Cai ZR, Chen ML, Simard JF, Linos E. Assessing biases in medical decisions via clinician and AI chatbot responses to patient vignettes. JAMA Network Open. 2023;6(10):e2338050–e2338050.

115. Ji H, Wang X, Sia CH, Yap J, Lim ST, Djohan AH, et al. Large language model comparisons between English and Chinese query performance for cardiovascular prevention. Communications Medicine. 2025;5(1):177.

116. Ajit-Roger E, Moise A, Peralta C, Orishchak O, Daniel SJ. Enhancing Multilingual Patient Education: ChatGPT's Accuracy and Readability for SSNHL Queries in English and Spanish. OTO open. 2024;8(4):e70048.

117. Bejan CA, Reed AM, Mikula M, Zhang S, Xu Y, Fabbri D, et al. Large language models improve the identification of emergency department visits for symptomatic kidney stones. Scientific Reports. 2025;15(1):3503.

118. Wan Z, Guo Y, Bao S, Wang Q, Malin BA. Evaluating Sex and Age Biases in Multimodal Large Language Models for Skin Disease Identification from Dermatoscopic Images. Health Data Science. 2025;5:0256.

119. Andreadis K, Newman DR, Twan C, Shunk A, Mann DM, Stevens ER. Mixed methods assessment of the influence of demographics on medical advice of ChatGPT. Journal of the American Medical Informatics Association. 2024;31(9):2002–2009.

120. Zhao Y, Wang H, Liu Y, Suhuang W, Wu X, Zheng Y. Can LLMs replace clinical doctors? Exploring bias in disease diagnosis by large language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024; 2024. p. 13914–13935.

121. Liu X, Wu J, Shao A, Shen W, Ye P, Wang Y, et al. Uncovering language disparity of ChatGPT on retinal vascular disease classification: cross-sectional study. Journal of medical Internet research. 2024;26:e51926.

122. Arslan B, Nuhoglu C, Satici M, Altinbilek E. Evaluating LLM-based generative AI tools in emergency triage: A comparative study of ChatGPT Plus, Copilot Pro, and triage nurses. The American Journal of Emergency Medicine. 2025;89:174–181.

123. Ayoub NF, Balakrishnan K, Ayoub MS, Barrett TF, David AP, Gray ST. Inherent bias in large language models: a random sampling analysis. Mayo Clinic Proceedings: Digital Health. 2024;2(2):186–191.

124. Guerra-Adames A, Avalos M, Dorémus O, Gil-Jardiné C, Lagarde E. Uncovering Judgment Biases in Emergency Triage: A Public Health Approach Based on Large Language Models. Proceedings for Machine Learning Research (PMLR). 2024;.

125. Chen S, Gallifant J, Gao M, Moreira P, Munch N, Muthukkumar A, et al. Cross-care: assessing the healthcare implications of pre-training data on language model bias. Advances in Neural Information Processing Systems. 2024;37:23756–23795.

126. O'Malley A, Veenhuizen M, Ahmed A. Ensuring appropriate representation in artificial intelligence-generated medical imagery: Protocol for a methodological approach to address skin tone bias. Jmir ai. 2024;3:e58275.

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

128. Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:231211805. 2023;.

129. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems. 2022;35:24824–24837.

130. Klingbeil A, Grützner C, Schreck P. Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior. 2024;160:108352.

131. Steyvers M, Tejeda H, Kumar A, Belem C, Karny S, Hu X, et al. What large language models know and what people think they know. Nature Machine Intelligence. 2025;7(2):221–231.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1Table.pdf
- S2Table.pdf
- S3Table.pdf