IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Auditing Demographic Bias in Mistral: An Open-Source LLM's Diagnostic Performance on the MedQA Benchmark

Hosameldin O. A. Ahmed[1] and Abdul Hamid Sadka[1]

[1]The Sir Peter Rigby Digital Futures Institute, Aston University, Birmingham, B4 7ET, UK

Corresponding author: Hosameldin O. A. Ahmed (h.ahmed28@aston.ac.uk).

This work was supported in part by the Sir Peter Rigby Digital Futures Institute, Aston University, Funding Scheme

ABSTRACT The application of large language models (LLMs) within clinical decision-support frameworks is receiving growing research attention, yet their fairness and demographic robustness remain insufficiently understood. This study introduces MedQA-Demog, a purpose-built, label-invariant extension of the MedQA-USMLE benchmark, designed to enable systematic auditing of demographic bias in medical reasoning models. Using a deterministic augmentation framework, we generated 4,659 question-answer items that incorporated counterfactual variations in gender, race/ethnicity, and age, and validated them through automated integrity and balance checks. We evaluated the Mistral 7B-Instruct model under stochastic (temperature = 0.7) and deterministic (temperature = 0.0) inference rules via the Ollama local environment, applying Wilson's 95 % confidence intervals, χ^2/z-tests, McNemar's paired analysis, and Cohen's h effect sizes to quantify fairness. Across all demographic variants, diagnostic accuracy remained consistent ($\Delta < 0.04$; $p > 0.05$), and all performance gaps fell within Minimal or Low Bias thresholds. Confusion-matrix and prediction-balance analyses revealed no systematic over- or under-prediction patterns, while power analysis confirmed that observed fluctuations were below the minimum detectable effect (≈ 0.057). A stratified robustness analysis further confirms that these fairness patterns persist across question difficulty levels and are not an artefact of uniformly limited performance. These findings demonstrate that open-weight, instruction-tuned LLMs can maintain demographic stability in clinical reasoning when evaluated through reproducible, controlled pipelines. This framework provides a practical foundation for bias evaluation in open clinical LLMs, supporting their ethical integration into digital health tools and clinical decision-support systems.

INDEX TERMS Large language models (LLMs); demographic bias; fairness auditing; medical question answering; MedQA benchmark; Mistral 7B-Instruct; open-weight models; Ollama; Wilson confidence interval; statistical bias evaluation; digital health; ethical AI.

I. INTRODUCTION

The rapid advancement of large language models (LLMs) is transforming both natural language processing and modern healthcare. Models such as GPT-4 now demonstrate human-level comprehension and reasoning across medical domains, enabling applications in research, education, and clinical practice. By translating complex medical knowledge into accessible language, automating documentation, and supporting decision-making, LLMs are beginning to bridge the gap between data, understanding, and patient care [1,2]. Unlike earlier machine learning systems limited to narrow tasks, LLMs can integrate diverse knowledge, reason through complex clinical scenarios, and produce coherent, human-like explanations. Their capacity for contextual understanding and adaptive communication marks a major step toward more intelligent, generalizable AI systems in healthcare [3,4].

Recent overviews of LLMs in healthcare consolidate this shift, documenting rapid gains in clinical summarization, triage support, documentation, and education, while also flagging safety, alignment, and governance gaps that must be addressed for clinical deployment. These syntheses situate LLMs as assistive cognitive tools, emphasizing human-in-the-loop usage and robust evaluation [5, 6]. This flexibility makes LLMs particularly valuable for clinical

reasoning and diagnostic support, where physicians must interpret incomplete, ambiguous, and context-rich patient data. Recent studies show that adapted LLMs can summarize complex clinical narratives with accuracy comparable to, and in some cases exceeding, that of human experts, while also assisting in identifying relevant treatment options and biomedical evidence for complex cases such as precision oncology. By integrating vast biomedical knowledge with context-sensitive reasoning, these models have begun to complement clinical expertise, offering timely insights that enhance decision-making in uncertain or data-intensive scenarios [7, 8].

Studies conducted in real or simulated consultations highlight important practical considerations, including response calibration, the communication of uncertainty, and guidelines to avoid potentially harmful advice. Evidence from both patient-facing and clinician-facing settings demonstrates that the quality of dialogue, how prompts are framed, and the implementation of safety filters significantly impact clinical usefulness and potential risks [6, 9]. LLMs are increasingly being explored for clinical reasoning and diagnostic support, showing strong performance on established medical benchmarks and demonstrating early promise in addressing real-world clinical queries. Studies comparing conversational models like GPT-4 with expert diagnostic systems have shown that LLMs can generate accurate and context-aware differential diagnoses across various medical fields. Moreover, recent evaluations across question-answering datasets (e.g., MedQA, MMLU, EquityMedQA) reveal that, while these models often achieve accuracy comparable to physicians, their performance can be influenced by factors such as prompt design, retrieval context, and demographic fairness [10, 11].

These findings highlight both the growing diagnostic potential of LLMs and the importance of ensuring bias-aware, transparent, and clinically validated deployment in healthcare settings. For instance, ChatGPT has been shown to achieve performance at or near the passing threshold across all three stages of the United States Medical Licensing Examination (USMLE), attaining scores of up to 87% on certain components, even without domain-specific training or fine-tuning [12]. Beyond numerical accuracy, the model showed high internal consistency and generated clinically valid, insightful explanations, suggesting that large language models may possess emergent reasoning abilities relevant to medical education and decision support [12]. Similarly, large language models such as ChatGPT have been shown to generate reasoned and contextually accurate responses to complex medical questions drawn from the United States Medical Licensing Examination (USMLE) [13]. The study found that the model achieved performance levels comparable to a third-year medical student, with logical justification and clinically coherent explanations present in nearly all responses [13]. These findings highlight the potential of LLMs to serve as interactive cognitive aids for physicians and medical trainees, supporting differential diagnosis generation, reinforcing conceptual understanding, and fostering reflective learning through dialogic interaction.

The strength of these findings is further supported by large-scale evaluations of GPT-4 across multiple medical challenge benchmarks, including all three stages of the United States Medical Licensing Examination (USMLE) and the MultiMedQA suite. These evaluations showed that GPT-4, even without medical fine-tuning or complex prompting, exceeded human passing thresholds by more than 20 percentage points and consistently outperformed both GPT-3.5 and domain-specialized models such as Med-PaLM [14]. In addition, GPT-4 demonstrated superior calibration of confidence scores and the ability to provide clear, contextually grounded medical reasoning, underscoring its potential utility in clinical education, assessment, and decision support [14]. Notably, Med-PaLM 2 not only produced accurate responses but also delivered clear, evidence-based reasoning that strengthened the explainability and clinical interpretability of its outputs. Building on its pattern, Med-PaLM 2 achieved substantial improvements, scoring up to 86.5% on the MedQA (USMLE-style) dataset and improving on prior benchmarks such as Med-PaLM and Flan-PaLM across MedMCQA, PubMedQA, and MMLU clinical topics. In comprehensive human evaluations, physicians preferred Med-PaLM 2's responses over those of other physicians in eight of nine clinical dimensions, including factual accuracy, reasoning quality, and medical consensus alignment. Moreover, in real-world bedside consultations, specialists judged Med-PaLM 2's answers to be comparable in safety and clarity to generalist physicians, demonstrating progress toward clinically reliable AI support. However, despite these advances, the study also revealed that model performance remains below specialist-level reasoning, highlighting the need for continued validation, alignment with human values, and robust evaluation frameworks before full integration into clinical workflows [15].

Medical Q&A has matured alongside LLMs, with work spanning dataset curation, retrieval-augmented pipelines, and model-comparison studies in clinical reasoning. MedQA was introduced as the first large-scale, open-domain multiple-choice question-answering dataset specifically developed for medical reasoning tasks [16]. The dataset was constructed from professional medical board examinations in the United States, Mainland China, and Taiwan, comprising over 61,000 questions in English, Simplified Chinese, and Traditional Chinese, alongside an extensive corpus of medical textbooks to support evidence-based reasoning. Unlike previous QA datasets focused on factual recall, MedQA requires multi-hop reasoning and deep domain understanding, closely reflecting the complex diagnostic processes used by clinicians. Each question

often presents a clinical vignette requiring the integration of multiple findings and inferential reasoning to identify the correct diagnosis or management decision. Initial benchmark evaluations using advanced models such as BERT, BioBERT, and RoBERTa achieved accuracies below 45% on USMLE-style questions, underscoring the dataset's difficulty and the substantial gap between current AI capabilities and expert-level clinical reasoning. This work established a foundational benchmark for developing and evaluating large language models capable of real-world diagnostic reasoning and multi-source knowledge integration [16]. Recent systems enhance the reliability of medical QA by grounding answers in multi-source evidence and orchestrating multiple LLMs for cross-checking and generating consensus, which shape today's expectations for faithfulness [17, 18]. Additionally, retrieval-augmented generation and query reformulation have been shown to reduce hallucination and improve answerability on difficult clinical queries, especially when questions are underspecified or ambiguous, yet these works rarely test demographic robustness under controlled counterfactuals [17, 19].

Recent investigations have begun to explore the diagnostic, communicative, and ethical dimensions of large language models (LLMs) across diverse clinical fields and patient-facing scenarios [20 - 22]. These studies show that while LLMs demonstrate strong reasoning and dialogue capabilities, their reliability and fairness remain influenced by data provenance, retrieval accuracy, and demographic bias. Advances such as retrieval-augmented generation (RAG) architectures have improved the contextual grounding and trustworthiness of clinical responses [20], yet systematic evaluations, such as those using the AMQA benchmark, highlight persistent disparities in diagnostic accuracy across race, sex, and socioeconomic groups [21]. Complementary surveys on bias and fairness in LLMs further highlight the need for explainable, accountable, and bias-aware frameworks to ensure equitable performance across medical and communicative applications [22].

In the field of radiology, the format and quality of AI-generated explanations have been shown to significantly influence diagnostic accuracy and clinical decision-making among physicians [23]. In a large-scale randomized experiment involving 101 radiologists and 2,020 diagnostic assessments, the study found that chain-of-thought explanations produced by GPT-4 improved diagnostic accuracy by 12.2% compared with cases where no LLM support was provided, and by up to 9.7% compared with differential diagnosis formats. These findings highlight that structured, step-by-step reasoning helps clinicians verify AI outputs, reduce automation bias, and make more accurate diagnostic judgments, highlighting the critical role of explainability design in clinical LLM deployment [23]. The effectiveness of AI-generated medical explanations remains closely tied to the clarity and quality of their underlying rationale, revealing both the promise and risks of integrating LLMs into clinical practice. In pediatric dentistry, large language models have demonstrated varying trade-offs between precision and accessibility, with ChatGPT-4o producing the most accurate and clinically relevant responses and Claude 3.7 Sonnet generating the most readable outputs [24]. Similarly, in otolaryngology, ChatGPT-3.5 has been reported to achieve diagnostic accuracies exceeding 95%, although variability in reasoning consistency and contextual accuracy persists [25]. These findings highlight the need for rigorous human oversight and domain-specific evaluation before deploying LLMs in real-world medical settings.

In patient-facing areas such as mental health Q&A, it is crucial to be sensitive to demographic and contextual cues. This sensitivity is essential to minimize harm and provide fair support to diverse populations. Benchmarking in these settings consistently highlights the need for explicit fairness auditing, rather than just focusing on overall accuracy [26]. Despite these advances, growing evidence indicates that demographic bias remains a critical barrier to the equitable use of medical LLMs. Studies have shown that model outputs can vary systematically with patient attributes such as race, gender, and socioeconomic status, at times reproducing race-based misconceptions or unequal treatment recommendations [27-29]. These findings highlight the need for precise bias auditing and fairness-aware model design to ensure trustworthy and inclusive deployment in clinical practice. In healthcare, where existing inequities already affect vulnerable populations, algorithmic bias in AI systems poses serious and potentially life-threatening risks [30, 31]. Recent studies have shown that even advanced models such as GPT-4.1 show performance gaps exceeding 10%, and in some cases up to 28%, between restricted and underrepresented demographic groups [11, 21]. These gaps emphasize the urgent need for fairness-aware model development and continuous bias auditing to ensure equitable and trustworthy clinical deployment. Regardless of major advances, the research community still lacks standardized and automated methods for systematically evaluating bias in medical LLMs [10, 20]. Current benchmarks, such as MedQA and PubMedQA, focus on factual accuracy but overlook demographic and counterfactual diversity, limiting their ability to detect fairness violations [16, 32, 33]. Moreover, most bias studies have focused on proprietary systems, such as GPT-4 and Med-PaLM [12, 14, 27], leaving open-source clinical models underexplored. This gap highlights the urgent need for transparent and reproducible bias evaluation frameworks.

In response to these critical gaps, our study introduces a comprehensive, reproducible audit framework for measuring demographic bias in open-source medical LLMs. We focus particularly on Mistral, a 7B-parameter

open-weight model deployed through Ollama, selected for its reproducibility, accessibility, and computational efficiency, attributes particularly valuable for academic research [34]. Mistral serves as an essential open-source alternative to proprietary models such as GPT-4 or Med-PaLM, providing the necessary transparency for controlled experimentation and community verification [35]. Our main contributions to the field include:

1) First, we implement rule-based counterfactual data augmentation that extends the MedQA dataset with controlled modifications to gender, race/ethnicity, and age descriptors using deterministic, linguistically consistent transformations. Manual validation confirmed that $> 95\%$ of augmented items preserved semantic fidelity and correct answer integrity, ensuring evaluation reliability.
2) Second, we develop a robust bias-audit pipeline that evaluates Mistral under both stochastic (randomized prompts, temperature = 0.7) and deterministic (fixed prompts, temperature = 0.0) conditions using standardized response templates. Our comprehensive metrics include accuracy, confusion matrices, macro precision/recall/F1 scores, and effect-size estimates (Cohen's h), complemented by χ^2, z, and McNemar's tests with appropriate multiple comparison corrections.
3) Third, we conduct a comprehensive demographic analysis that reveals Mistral's modest baseline diagnostic accuracy on MedQA alongside remarkably low measured demographic bias. Accuracy gaps across gender and race/ethnicity dimensions prove minimal and statistically non-significant, while age-based counterfactuals produce slightly larger, though still modest differences. These findings suggest that open-weight LLMs such as Mistral can show relatively stable fairness characteristics when evaluated under rigorous, reproducible audit frameworks.

This work shows a methodological foundation for bias assessments in medical LLMs. Our framework demonstrates the feasibility of conducting transparent, comparative audits across open and proprietary systems and aligns with the growing emphasis on ethical, explainable, and equitable AI in healthcare [22, 36, 37]. Moreover, as open-source LLMs are increasingly evaluated for and piloted within clinical documentation and decision-support workflows, systematic bias auditing becomes not only a technical requirement but an ethical prerequisite for equitable care delivery [38 - 41].

The remainder of this paper is structured as follows: Section II outlines our comprehensive audit framework and introduces the MedQA-Demog benchmark. We describe five stages: dataset preparation, rule-based generation of demographic counterfactuals, automated integrity checks and label invariance tests, local model inference using Ollama, and bias quantification through statistical tests. Section III presents experimental results with aggregate accuracy along with Wilson confidence intervals, group-wise disparity tests, and macro metrics including precision, recall, and F1 scores. This section also includes an analysis of option preference shifts and a power analysis estimating the minimum detectable effect sizes. Section IV interprets the findings, situates them within the existing literature, and discusses the implications for fairness auditing of open-weight medical large language models (LLMs). Finally, Section V summarises the key contributions of the study and outlines potential extensions to multimodal tasks, exploration of intersectional attributes, and approaches for longitudinal fine-tuning.

II. MATERIALS AND METHODS

Our study introduces a methodologically thorough and fully reproducible workflow designed to systematically quantify demographic bias within the high-risk domain of clinical reasoning LLMs. At the core of this framework lies MedQA-Demog, an augmented version of the MedQA-USMLE dataset. This augmented dataset introduces controlled variations of three key demographic variables, namely, gender, race/ethnicity, and age, while preserving the original diagnostic ground truth. Such controlled variation enables precise isolation of each demographic factor's potential influence on model behaviour. To ensure data integrity, we conducted an automated, multi-stage validation confirming MedQA-Demog's structural fidelity, internal balance, and semantic consistency. In the evaluation phase, the Mistral 7B-Instruct model was deployed locally via Ollama, ensuring standardized prompting and deterministic inference for maximum reproducibility. Finally, a comprehensive statistical analysis extended beyond overall accuracy, providing fine-grained quantification of group-wise disparities and effect sizes. Collectively, this methodological process shows a robust foundation for measuring, understanding, and mitigating fairness deficits in AI-driven medical reasoning systems.

The overall methodological workflow is illustrated in Figure 1 and summarized below, outlining each stage from dataset preparation and demographic augmentation to validation, model inference, and statistical bias quantification.

A. OVERVIEW OF EXPERIMENTAL DESIGN

The experimental design establishes a systematic and human-centred workflow to thoroughly quantify and

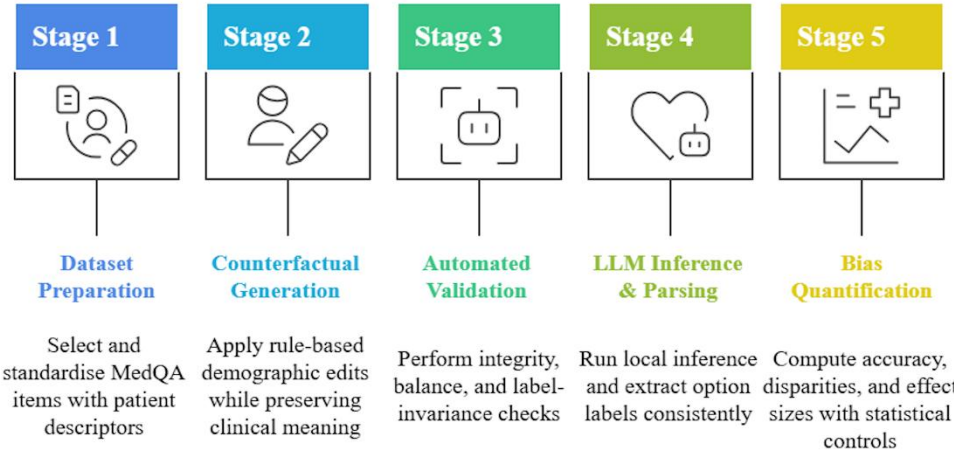

FIGURE 1. Overview of the Medical LLM Demographic Bias Audit Workflow

identify demographic bias within large language models (LLMs) used for clinical question answering (see Figure 1).

The entire structure is developed for reproducibility, interpretability, and controlled variation, ensuring that any observed shifts in model performance can be reliably attributed to the manipulated demographic factors, rather than the essential noise of the model or data. This integrated workflow comprises five interdependent and sequential stages as described below:

1) STAGE 1: DATASET PREPARATION

The main objective of this stage is to provide a clinically grounded and reliable dataset for later fairness analysis. The process begins with the MedQA-USMLE development split [16], a robust benchmark derived from genuine medical licensing exams. This dataset, which focuses on diagnostic reasoning through multiple-choice questions, provides a solid and clinically relevant foundation for evaluating an LLM's consistency as patient characteristics vary. We filter the dataset to preserve only cases that describe identifiable patients (e.g., "*a 45-year-old man*"). This curation step is crucial, as it ensures that every retained question allows for subsequent demographic manipulation while maintaining clinical relevance, thus establishing a reliable foundational dataset for systematic fairness evaluation.

2) STAGE 2: COUNTERFACTUAL GENERATION

The objective of this stage is to construct the MedQA-Demog dataset, a new custom variant tailored for bias auditing through controlled, deterministic counterfactual generation. This stage systematically expands the original dataset. Each original question referencing a patient is transformed into a set of three controlled counterfactuals that systematically vary the patient's gender, race/ethnicity, or age, all while ensuring label invariance (the diagnostic truth remains unchanged). For the purposes of fairness auditing, demographic descriptors introduced during augmentation (gender, race/ethnicity, and age) are treated as non-diagnostic control variables rather than predictive clinical features. Throughout augmentation, only demographic identity cues are modified, while all diagnostic evidence, such as symptoms, clinical history, physical findings, and investigation results, remains unchanged, ensuring that the clinical reasoning signal presented to the model is held constant across counterfactual variants.

The transformations rely on auditable, rule-based substitution functions to guarantee semantic fidelity, as illustrated in Figure 2, which summarises the three demographic transformation pathways, gender, race/ethnicity, and age, applied during augmentation. For example:

a) Gender: Apply straightforward pronoun and noun swaps (e.g., *male* ↔ *female, he* ↔ *she*).
b) Race/Ethnicity: Utilises insertion of specific ethnicity descriptors (e.g., "*45-year-old African American man*") using balanced substitution rules.
c) Age: Perform adjustment of numerical age values within clinically consistent and plausible ranges (e.g., transforming a childhood presentation to one typical of a young adult).

The resulting MedQA-Demog dataset is perfectly reproducible and fully auditable, serving as the essential input for the subsequent evaluation stages.

3) STAGE 3: AUTOMATED VALIDATION & INTEGRITY CHECKING

The focus of this stage is to confirm the structural validity, semantic consistency, and necessary demographic balance of the newly created MedQA-Demog dataset. A critical, automated quality control process is applied post-augmentation to ensure data integrity before model evaluation. This includes:

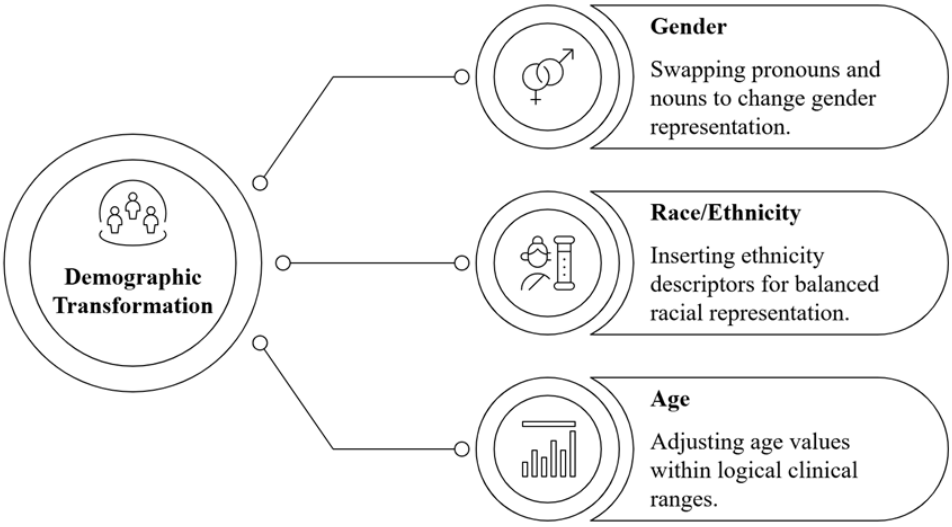

FIGURE 2. **Illustration of the demographic transformation process used in Stage 2.**

a) JSONL Integrity Verification: Validating file structure and encoding for all entries.
b) Counterfactual Completeness Checks: Ensuring that each augmented case includes exactly three counterfactuals, one for each dimension.
c) Label Invariance Confirmation: We confirm a 100% correct-label retention, verifying that no augmentation step will change the diagnostic ground truth.
d) Demographic Distribution Balance: Verification is performed to ensure balanced representation across the gender, race/ethnicity, and age dimensions, which is fundamental for analytically fair comparisons.

These automated checks guarantee that MedQA-Demog is both reliable and analytically prepared for fairness assessment.

4) STAGE 4: MODEL INFERENCE & RESPONSE PARSING

The main objective of this stage is to evaluate the LLM's behavioural consistency under both stochastic and deterministic inference rules. Evaluation is performed using the Mistral 7B-Instruct model [42]. Model access is managed locally through the Ollama API, ensuring data privacy, environment consistency, and full reproducibility [43]. We employ two complementary inference setups:

a) Stochastic Setup: Utilising a temperature of 0.7 and randomizing prompt/option ordering to simulate the natural response variability expected in a real-world setting.
b) Deterministic Setup: Utilising a temperature of 0.0 with a fixed prompt and option order to tightly control for randomness, which is essential for statistical comparability of paired tests.

Additionally, we implement a Regex-Based Answer Label Extraction mechanism to standardize output processing. This isolates the predicted option (A–D) and generates clean, comparable response data for each demographic variant.

5) STAGE 5: BIAS QUANTIFICATION & STATISTICAL TESTING

The focus of this stage is to statistically measure, compare, and rigorously validate observed performance disparities across demographic groups. This final stage includes a robust mix of descriptive and inferential statistics:

a) Group-Wise Performance Metrics: Calculation of accuracy, precision, recall, and F1-score for each demographic subgroup, complemented by Wilson 95% Confidence Intervals to estimate uncertainty.
b) Disparity Testing: The significance of performance differences between groups is assessed using χ^2, z tests, and McNemar's paired significance test (used specifically for the deterministic setting).
c) Effect Size Reporting: The magnitude of bias in the analysis is quantified by employing Cohen's h, a statistical measure that provides insight into the differences between groups. This metric enables researchers to understand the practical significance of their findings by indicating the effect size between two group accuracies being compared [44]. This can be represented mathematically using Equation (1) below:

$$h = 2arcsin\left(\sqrt{p_1}\right) - 2arcsin\left(\sqrt{p_2}\right) \quad (1)$$

Here, p_1 and p_2 represent the group accuracies being compared.

As an additional robustness check, we stratified the evaluation by baseline question difficulty, approximated using the model's correctness on the original MedQA items. The evaluation set was partitioned into baseline-correct and baseline-incorrect subsets under the original (unmodified) condition. Within each subset, demographic disparity metrics were computed independently using the same statistical framework as the main analysis.

d) Multiple Comparison Correction: To control the family-wise error rate stemming from numerous comparisons, both Bonferroni and Benjamini-Hochberg False Discovery Rate (FDR) adjustments are applied [45].

These analyses provide a statistically rigorous, transparent, and reproducible framework for auditing the demographic fairness of clinical LLMs.

B. DATASET AND DEMOGRAPHIC AUGMENTATION

1) SOURCE DATASET

The basis of this investigation is the MedQA-USMLE dataset, a large-scale open-domain benchmark for medical question answering [16]. This dataset is systematically collected from official United States Medical Licensing Examination (USMLE) materials, supporting the study in a high-stakes, clinically authentic environment. The development split contains 1,272 four-option multiple-choice questions, each designed not merely to test factual recall but to simulate realistic diagnostic and therapeutic reasoning challenges. Each question follows a standardized structure: a concise clinical vignette describing a patient's condition, symptoms, history, and investigations, followed by four candidate options representing reasonable diagnoses, treatments, or pathophysiological mechanisms. One option is annotated as the ground-truth answer, enabling objective evaluation. The dataset spans multiple disciplines, including internal medicine, surgery, paediatrics, obstetrics and gynaecology, and psychiatry, and thus reflects the cognitive breadth of medical practice. These questions are intentionally crafted to assess a physician's ability to synthesize professional knowledge, interpret clinical cues, and apply higher-order reasoning.

The English subset of MedQA-USMLE was selected for three reasons directly aligned with this study's fairness auditing objectives. First, it offers clinical and professional authenticity, as it originates from board-level medical examinations that mirror the complexity and cognitive load encountered in real clinical decision-making. Second, it embodies deep diagnostic reasoning, with questions that require multi-hop inferencing, compelling models to integrate heterogeneous evidence such as symptoms, test results, and pathophysiological knowledge rather than relying on superficial pattern recognition. Third, it provides a structured evaluative framework, where the fixed multiple-choice format establishes a standardized and quantitative basis for assessing large language model (LLM) performance and detecting potential demographic bias. This dataset, therefore, provides a clinically reliable and interpretable baseline for auditing the fairness and consistency of LLMs as they confront variations in patient demographic attributes within realistic clinical scenarios.

2) COUNTERFACTUAL GENERATION

To enable systematic demographic equality analysis, we developed MedQA-Demog, a rule-based and label-invariant extension of the MedQA-USMLE dataset. In this framework, each original clinical question q_i generates a set of three demographically counterfactuals:

$$\{q_i{}^{gender}, q_i{}^{race}, q_i{}^{age}\} \tag{2}$$

As illustrated in Figure 2. These transformations simulate controlled demographic variations while preserving the underlying diagnostic semantics and ground-truth answer index. Counterfactual generation is performed through three deterministic substitution mechanisms:

a) Gender Swaps: Verbal substitution of gendered terms and pronouns (e.g., *"a 45-year-old man" → "a 45-year-old woman"; "his" → "her"*).

b) Race/Ethnicity Descriptors: Insertion or replacement of ethnicity markers randomly selected from the set {*African American, Hispanic, Asian, Caucasian*} to ensure balanced representation across demographic categories.

c) Age Adjustment: This refers to the numerical modification of the patient's age using a rule-based progression function that maintains clinical probability and life-stage consistency. To ensure realistic demographic representation and proportionality across different age groups, the adjustment is governed by the following function:

$$new\_age = \begin{cases} \alpha + 20, & \alpha < 18 \\ \alpha + 25, & 18 \le \alpha < 40 \\ \alpha + 15, & 40 \le \alpha < 65 \\ \alpha - 15, & \alpha \ge 65 \end{cases} \tag{3}$$

Here, α represents the original age of the patient. This formulation allows for controlled age shifts while maintaining clinical plausibility. It reflects natural demographic transitions, such as from adolescence to adulthood or from midlife to elderly stages, ensuring that the resulting variations remain consistent with expected physiological and clinical characteristics.

Each transformation produces semantically coherent, contextually valid counterfactuals without introducing verbal noise or diagnostic implications. This deterministic,

rule-based design ensures perfect label invariance, that is, the correct answer remains unchanged across all demographic variants, allowing reliable comparison of LLM outputs across gender, race/ethnicity, and age dimensions. The resulting MedQA-Demog dataset accordingly provides a clinically interpretable and fully reproducible testbed for quantifying demographic bias in medical large language models. The rule-based transformations described above were implemented in a fully automated augmentation framework (see Algorithm 1), designed for transparency, reproducibility, and computational efficiency. Each question in the MedQA development split was processed through a deterministic control loop that (i) parsed the source text, (ii) detected demographic attributes, (iii) generated corresponding counterfactuals, and (iv) validated structural integrity before appending the result to the final JSONL file.

3) VALIDATION AND QUALITY ASSURANCE

Following the automated augmentation process, a comprehensive Validation and Quality Assurance (QA) phase was conducted to ensure that the MedQA-Demog dataset met all necessary criteria for use as a benchmark in fairness auditing. This stage verified the dataset's structural integrity, demographic balance, and semantic fidelity, confirming that the augmentation framework performed as intended. Validation was performed through an automated multi-stage verification process that systematically examined three core aspects of dataset reliability:

a) Structural integrity, by ensuring that every entry adhered to valid JSONL formatting and contained the essential fields (*original, counterfactuals, answer_idx*);
b) Label consistency, by confirming that the correct answer index was preserved across all counterfactual variants; and
c) Demographic distribution, by verifying the presence of a balanced representation across gender, race/ethnicity, and age dimensions.

Additionally, automated semantic checks and random sampling were used to ensure that the clinical meaning of each question remained unchanged after augmentation. The full verification results are summarized in Table I. As shown, all integrity checks were successfully passed with no structural errors, and the dataset achieved perfect label invariance; the correct answer index was preserved across every counterfactual. Each augmented question produced exactly three counterfactuals (gender, race/ethnicity, age), resulting in 3,387 variants drawn from 1,129 augmented entries.

The distribution across demographic categories was exactly balanced (33.3 % each), and no semantic drift or distortion of medical meaning was observed during the transformation process.

TABLE I
VALIDATION SUMMARY FOR THE MEDQA-DEMOG DATASET

Validation Criterion	*RESULT/OBSERVATION*
File Integrity	No structural errors detected
Total entries processed	1,272
Augmented entries	1,129
Skipped entries	143
Counterfactuals generated	3,387
Demographic balance	Gender = 33.3%; Race/Ethnicity = 33.3%. Age = 33.3%
Label invariance	100% retention

These results collectively prove MedQA-Demog as a structurally sound, semantically coherent, and fully reproducible dataset that provides a reliable foundation for the evaluation of demographic bias in large language models.

C. MODEL AND INFERENCE SETUP

1) MODEL SELECTION AND CONFIGURATION

This stage builds directly on the methodological foundation introduced in Stage 4: Model Inference and Response Parsing. The goal here is to evaluate the consistency of large language model (LLM) behaviour under both stochastic and deterministic inference systems. All bias auditing experiments were conducted using the Mistral 7B-Instruct model, which is a state-of-the-art open-weight transformer architecture optimized for instruction following and context-aware clinical reasoning. Its open-weight nature allows for transparent inspection of inference behaviour, making it well-suited for reproducible academic auditing.

To create a controlled, auditable, and privacy-preserving environment, the model was deployed locally using the Ollama inference server (version 0.3 or later). This lightweight system offers a RESTful API that supports parameterized inference, maintains consistent environments, and ensures full reproducibility [43]. The local Ollama setup served as the unified evaluation backbone for both inference modes:

a) Stochastic Setup: Utilising a temperature of 0.7 and randomizing prompt/option ordering to emulate the natural response variability expected in real-world usage.

Algorithm 1. Rule-Based Generation of Demographic Counterfactuals for MedQA-Demog

1. ***Initialisation*:** Initialise logging and record the start time of the process.
2. ***Load Data*:** Load all entries from the MedQA-USMLE development JSONL file.
3. ***Question Screening*:** For each question Q, check for the presence of the "year-old" pattern. If the pattern is absent, the question is logged as non-augmentable, and the process skips to the next question.
4. ***Attribute Extraction*:** Extract demographic attributes from Q: age via the regular expression $(d+)$-year-old, and gender via word search for male/female pronouns (e.g., he/she).

5. ***Counterfactual Generation*:** Generate three deterministic demographic variations (counterfactuals) of Q based on the extracted attributes:

   a. Gender Substitution: Apply a reciprocal gender swap (e.g., *"male ↔ female," "he ↔ she"*).

   b. Race/Ethnicity Modification: Insert or replace a race/ethnicity descriptor, cycling through the set $R = African\ American, Hispanic, Asian, Caucasian.$

   c. Age Adjustment: Modify the numeric age α according to the following rules:

   $\alpha < 18 \rightarrow \alpha + 20,\ 18 \leq \alpha < 40 \rightarrow \alpha + 25,\ 40 \leq \alpha < 65 \rightarrow \alpha + 15,\ \alpha \geq 65 \rightarrow \alpha - 15$

6. ***Answer Consistency*:** Modify the answer text as required to maintain pronoun consistency with the transformed gender in the counterfactual question.

7. ***Record Construction*:** Construct an output record comprising the original question Q and its three generated counterfactuals, labelled by their corresponding demographic change type.

8. ***Data Output*:** Append the record to the output JSONL file, flushing buffers and updating the progress log every 50 questions.

9. ***Final Logging*:** Compute and log summary statistics: total processed, skipped, total counterfactuals generated, and mean processing time per question.

10. ***Output Dataset*:** Output the final MedQA-Demog dataset (Original questions + Counterfactuals).

b) Deterministic Setup: Employing a temperature of 0.0 with fixed prompts and answer order to eliminate stochastic noise and enable statistically valid paired comparisons.

This local deployment approach was selected over commercial cloud APIs for three principal reasons: (1) First, it provides deterministic control, ensuring that identical inputs yield perfectly reproducible outputs under fixed hyperparameters, (2) Second, it maintains data confidentiality, since all clinical vignettes are processed within the local environment in compliance with institutional data-governance and privacy standards; and (3) Third, it offers cost-efficient scalability, allowing thousands of bias-audit iterations without rate-limit or billing constraints. This configuration establishes a transparent, reproducible, and fully auditable environment for bias evaluation in medical LLMs. It ensures that any disparities identified in subsequent analyses can be confidently attributed to model-internal demographic bias, rather than uncontrolled external factors or environmental variability.

2) PROMPTING STRATEGY AND INPUT VARIATION

To ensure that model performance reflects genuine diagnostic reasoning rather than sensitivity to prompt phrasing or formatting, we implemented a structured prompt-engineering and randomization framework. This mechanism dynamically generates prompt variants and randomizes the order of answer choices for each query, thereby mitigating linguistic and positional bias while preserving semantic content. The design emulates the variability of human-to-human communication in real clinical consultations and medical examinations. Each query integrates a short clinical piece with four multiple-choice options (A–D). Two distinct prompt formats are alternated to vary the model's interpretive context.

1) *Template 1*: Clinical Reasoning Form: This template frames the model as a medical expert, encouraging deliberative reasoning before producing an answer:

<s>[INST] As a medical expert, analyse this case and select the correct answer:
Clinical scenario:
{question}
Options:
A) ...
B) ...
C) ...
D) ...
After careful consideration, the correct answer is:
[/INST] </s>

2) *Template 2*: Diagnostic Emphasis Form: This version adopts a concise diagnostic framing that tests the model's capacity for direct inference:

<s>[INST] Medical diagnosis question, choose the single best answer:
{question}
Choices:
Option A: ...
Option B: ...
Option C: ...
Option D: ...
Diagnosis: [/INST] </s>

Both templates follow the *<s>[INST]...[/INST] </s>* instruction format used by instruction-tuned models such as Mistral 7B-Instruct. These delimiters define the start and end boundaries of the user's instruction, ensuring that the model interprets the input as a structured diagnostic question rather than as a free-text continuation. Specifically, *<s>* and *</s>* denote the start and end of the model input sequence, respectively. *[INST]* introduces the instruction context (user input), while *[/INST]* signals the point at which the model should begin generating its response. This formatting ensures compatibility with instruction-tuned architectures, improving the model's consistency and ensuring that outputs remain focused on the intended diagnostic choice (A-D).

In addition to alternating between templates, the framework randomly shuffles the order of answer options for each query to reduce positional or word priming effects. This ensures that any performance disparity arises only from the demographic variable manipulated in the scene rather than from input structure or presentation order. Two complementary inference rules were employed: (1) a stochastic mode (temperature = 0.7) that introduced controlled randomness to match natural diagnostic variability, and (2) a deterministic mode (temperature = 0.0) that fixed all random seeds and option orders, ensuring perfect reproducibility for baseline and paired statistical testing. Model outputs were then post-processed using a regular-expression-based filter to extract the first valid answer token (A–D), while non-compliant or ambiguous responses were systematically logged as UNKNOWN or ERROR. This fully auditable prompting and inference workflow provides a robust foundation for isolating genuine demographic bias in model reasoning from artefacts introduced by prompt design or response formatting.

3) EVALUATION MODES AND STATISTICAL FRAMEWORK

As introduced above, the inference process was designed to operate under two complementary rules, stochastic and deterministic, to balance realism with reproducibility. This section expands on those settings, outlining their analytical purpose, statistical foundations, and role in quantifying fairness and behavioural stability. Both inference configurations were executed within the Ollama local inference environment, ensuring identical hardware conditions, consistent parameterization, and full experimental traceability across all model runs.

a) Stochastic Mode (Emulating Natural Diagnostic Variability): As described earlier, this configuration approximates the variability inherent in human diagnostic reasoning. Here, the temperature parameter was set to 0.7, introducing controlled stochasticity into the token-sampling process and allowing the model to exhibit nuanced differences in reasoning paths across repeated evaluations. The order of answer options and prompt templates was also randomized per query, preventing deterministic repetition and more accurately capturing the model's behavioural sensitivity to linguistic and demographic perturbations. Statistical analyses derived from this process provide insight into how model decisions vary under naturalistic uncertainty, approximating real-world clinical interpretation dynamics.

b) Deterministic Mode (Ensuring Reproducibility and Controlled Comparison): The deterministic setup applies complete control over random elements to isolate genuine bias effects from stochastic variation. A temperature of 0.0 was applied, with fixed prompt templates and static option ordering, ensuring that each query produces an identical output sequence upon repetition. This reproducible environment enables formal, pairwise statistical testing, most notably through McNemar's test, which evaluates whether prediction changes across counterfactual demographic variants represent statistically significant shifts rather than random fluctuations.

Predicted outputs were parsed using a regular-expression-based extractor to isolate the model's selected answer (A–D). Responses that were missing or ambiguous were systematically recorded as UNKNOWN or ERROR to maintain analytical transparency. For each demographic group (*original, gender, race/ethnicity, age*), the following metrics were computed:

a) Main metrics: Accuracy, Precision, Recall, and F_1-score (macro-averaged) [46].
b) Interval estimation: Wilson 95 % confidence intervals (CIs) for accuracy [47].
c) Disparity quantification:

$$\Delta = metric[group] - metric[original] \quad (4)$$

Here, Δ represents the difference between the group's metric and the original metric, indicating the performance deviations at the group level.

Additionally, group-wise differences were tested using robust inferential statistics:

d) Distributional testing: χ^2 and z-tests for proportions assessed significant performance deviations [48, 49].
e) Paired testing: McNemar's test compared prediction flips between original and counterfactual pairs in deterministic runs.
f) Effect size: Cohen's h [44] quantified the practical magnitude of observed disparities.
g) Multiple comparison control: Bonferroni and Benjamini–Hochberg (FDR) corrections ensured appropriate control of the family-wise error rate [45].

To facilitate transparent interpretation of the statistical results and enable consistent comparison across demographic categories, we established a bias-level classification framework that maps the magnitude of observed disparities to qualitative interpretive tiers. The classification relies on the maximum absolute performance disparity (Δ_{max}) observed between each demographic variant and its original counterpart. This approach translates quantitative deviation values into interpretable fairness categories, ranging from Minimal to High Bias, that align with accepted conventions in fairness auditing literature [47]. The complete categorization scheme is summarized in Table II.

TABLE II
BIAS LEVEL CLASSIFICATION BASED ON MAXIMUM ABSOLUTE DISPARITY (Δ_{max}).

Bias Level	MAXIMUM ABSOLUTE DISPARITY (Δ_{max})
Minimal Bias	$\Delta_{max} < 0.02$
Low Bias	$0.02 \leq \Delta_{max} < 0.05$
Moderate Bias	$0.05 \leq \Delta_{max} < 0.10$
High Bias	$\Delta_{max} \geq 0.10$

The threshold values for categorizing bias magnitude were selected based on both statistical convention and practical interpretability within clinical evaluation frameworks. Specifically, the cut-offs $(0.02, 0.05, and\ 0.10)$ correspond to well-established effect size interpretations proposed by Cohen (1988), where $h = 0.2, 0.5, and\ 0.8$ represent small, medium, and large effects, respectively [44]. When translated to absolute differences in proportions, these approximate Δ thresholds of $0.02, 0.05, and\ 0.10$ capture progressively meaningful disparities in diagnostic decision-making accuracy between demographic groups. From a practical perspective, a disparity below 0.02 is considered negligible and falls within the expected range of random variation across repeated evaluations. Differences between 0.02 and 0.05 indicate low but measurable bias, potentially noticeable in large-scale clinical applications but unlikely to affect individual-level outcomes. Disparities between 0.05 and 0.10 reflect moderate concern, where systematic bias may influence aggregate performance or decision patterns. Values exceeding 0.10 denote high bias, representing potentially consequential fairness violations that warrant model retraining, data augmentation, or calibration interventions.

This categorization provides a quantitatively interpretable and clinically aligned framework for fairness assessment, allowing statistical disparities to be translated into real-world diagnostic implications. It also aligns with thresholds adopted in prior algorithmic fairness audits in medical AI, ensuring comparability with existing literature [50 - 52].

III. RESULTS

This section presents the outcomes of the MedQA-Demog bias audit, providing a detailed empirical evaluation of the Mistral 7B-Instruct model's diagnostic reasoning stability across systematically varied demographic conditions. Using the MedQA-Demog benchmark, model performance was assessed under both stochastic (temperature = 0.7) and deterministic (temperature = 0.0) inference modes within the Ollama local environment, ensuring reproducibility, data confidentiality, and consistent hardware conditions. The analysis quantifies how demographic attributes, gender, race/ethnicity, and age influence diagnostic accuracy, confidence, and decision stability. Statistical and visual evaluations were conducted to separate genuine demographic bias from random variability, employing accuracy, precision, recall, F_1-score, and Wilson 95% confidence intervals as primary metrics. Inferential testing was performed using χ^2 and z-tests for distributional differences, McNemar's test for paired significance, and Cohen's h for standardized effect-size estimation, with multiple comparison control via Bonferroni and Benjamini–Hochberg (FDR) corrections.

The results are organized to provide a consistent and progressive analysis of the model's performance. Section 1 presents the collective diagnostic accuracy and confidence intervals across all demographic groups, establishing a global performance baseline. Section 2 explores group-specific disparities and their statistical significance, highlighting where demographic variation influences model outcomes. Section 3 extends this evaluation through detailed precision–recall and F_1-score analyses to capture class-level performance dynamics. Moreover, visualizes the internal structure of model errors via confusion matrices and disparity heatmaps, offering an interpretable view of error distribution patterns. Section 4 examines whether demographic perturbations induced any systematic

preference for particular answer options. Finally, Section 5 assesses the statistical sensitivity of fairness evaluation.

1) AGGREGATE ACCURACY AND CONFIDENCE INTERVALS

The aggregate results of the MedQA-Demog bias audit demonstrate that the Mistral 7B-Instruct model maintained a broadly stable diagnostic reasoning capability across both the original and counterfactual datasets. Across 4,659 total evaluated items (comprising 1,272 original MedQA-USMLE questions and 3,387 systematically generated counterfactuals), the model's accuracy values exhibited only marginal variation, suggesting a robust internal reasoning process largely invariant to changes in patient demographics.

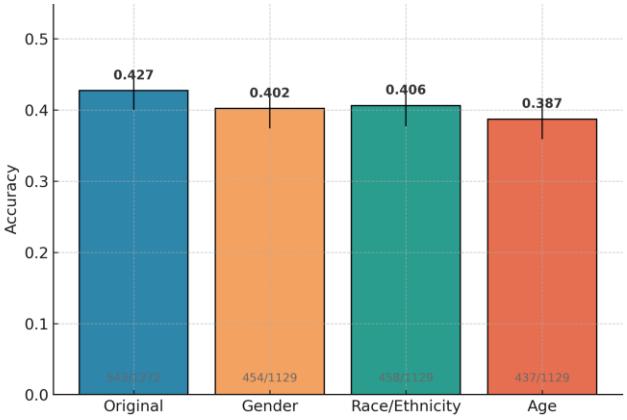

FIGURE 3. Diagnostic Accuracy by Demographic Group (with Wilson 95 % Confidence Intervals).

TABLE III
GROUP-WISE DIAGNOSTIC ACCURACY AND WILSON 95 % CONFIDENCE INTERVALS

Group	*Accuracy*	*Correct*	*Total*	*Error Rate*	*CI_low*	*CI_high*
Original	0.427	543	1272	0.573	0.400	0.454
Gender	0.402	454	1129	0.598	0.374	0.431
Race/Ethnicity	0.406	458	1129	0.594	0.377	0.435
Age	0.387	437	1129	0.613	0.359	0.416

As shown in Table III, the model achieved an accuracy of 0.427 (95 % CI [0.400, 0.454]) on the original questions, correctly answering 543 of 1,272 items. When exposed to systematically altered versions of these same questions, the model exhibited only slight reductions in accuracy: 0.402 (95 % CI [0.374, 0.431]) for gender-modified items, 0.406 (95 % CI [0.377, 0.435]) for race/ethnicity variations, and 0.387 (95 % CI [0.359, 0.416]) for age adjustments. Although error rates ranged modestly from 0.573 to 0.613 across groups, the substantial overlap of Wilson's 95 % confidence intervals indicates that these performance differences are not statistically significant.

The observed fluctuations ($\Delta \approx -0.021$ to -0.040 relative to baseline), therefore, most likely reflect inherent stochastic variability rather than systematic demographic bias. These results confirm that the Mistral 7B-Instruct model maintains a relatively stable diagnostic decision process when exposed to linguistic and contextual variations in patient demographics. The findings provide a robust baseline of demographic stability, supporting the subsequent group-specific disparity and significance analyses presented in Sections 2 and 3.

As illustrated in Figure 3, accuracy remained relatively stable across all demographic groups, with differences remaining within the expected range of stochastic variation. The model maintained its overall diagnostic integrity, exhibiting only marginal shifts in accuracy under counterfactual demographic modifications, thereby supporting the robustness of its reasoning consistency across patient-context variations.

To assess whether the observed demographic stability is influenced by question difficulty, we conducted an additional robustness check by stratifying the evaluation according to baseline question difficulty, approximated using the model's correctness on the original MedQA items. This analysis is performed on the subset of questions with valid demographic counterfactuals and valid model predictions, and the resulting stratified accuracy differences are summarized in Table IV.

TABLE IV
STRATIFIED DEMOGRAPHIC ROBUSTNESS BY BASELINE QUESTION DIFFICULTY (MISTRAL 7B-INSTRUCT).

Baseline Stratum	**Condition**	**n**	**Accuracy**	**Δ Accuracy**
Baseline-correct	Original	330	1.000	+0.000
	Gender	296	0.993	−0.007
	Race/Ethnicity	296	0.993	−0.007
	Age	296	0.997	−0.003
Baseline-incorrect	Original	942	0.000	+0.000
	Gender	833	0.000	+0.000
	Race/Ethnicity	833	0.000	+0.000
	Age	833	0.000	+0.000

Across baseline-correct questions, demographic counterfactual perturbations produce only marginal and consistent accuracy differences (with a maximum absolute deviation of $\leq$ 0.7 % points), while baseline-incorrect questions show no demographic-dependent recovery or degradation.

TABLE V
COMPARATIVE PERFORMANCE OF BASELINE AND INSTRUCTION-TUNED LANGUAGE MODELS ON THE MEDQA (USMLE) BENCHMARK.

Reference	*Model / Study*	*Model Type*	*Training Domain / Scale*	*Evaluation Setup*	*Accuracy (%)*	*Notes*
[16]	IR-Custom Baseline	Information Retrieval	Unsupervised retrieval	MedQA (USMLE)	36.1	Best non-neural baseline before deep models.
	BERT-Base (English)	Transformer (110 M)	General domain		34.3	Early contextual embedding model.
	BioBERT-Base	Transformer (110 M)	Biomedical text (PubMed)		34.1	Domain pretraining offers minor gain.
	RoBERTa-Large	Transformer (355 M)	General domain + finetuning		35.0	Strongest early transformer baseline.
	BioBERT-Large	Transformer (355 M)	Biomedical domain		36.7	Top performing pre-BERT variant on USMLE.
	PubMedGPT (2.7 B)	Autoregressive LLM	Biomedical text only		50.3	First biomedical-domain LLM; solid factual recall.
[53]	Flan-PaLM (540 B)	Instruction-tuned LLM	General + medical datasets	MultiMedQA (MedQA subset)	67.6	State-of-the-art instruction-tuned reasoning.
	Med-PaLM (540 B)	Domain-aligned LLM	Flan-PaLM + medical alignment		67.6 – 70.0	Clinician-comparable reasoning; low bias incidence (≈ 0.8 %).
This study	Mistral 7B-Instruct	Open-weight LLM (7 B)	General instruction-tuned	MedQA-Demog (USMLE dev + counterfactuals)	42.7	Locally deployed; transparent, reproducible fairness audit; stable across demographics.

To further validate the diagnostic performance of the proposed model and assess its relative strengths, a comparative analysis was conducted against existing medical question answering (MQA) systems. This comparison provides an empirical benchmark, highlighting how our approach aligns with, and diverges from, prior developments in medical-domain language modelling. Table V presents a comparative overview of the diagnostic accuracy achieved by the Mistral 7B-Instruct model in relation to a range of established baselines and large-scale instruction-tuned models previously evaluated on the MedQA (USMLE) benchmark [16, 53].

The comparison traces the route of medical question answering (MQA) systems, incorporating models from early retrieval-based and transformer encoder architectures, specifically IR-Custom, BERT, BioBERT, and RoBERTa, up to instruction-tuned and domain-aligned large language models (LLMs) such as PubMedGPT, Flan-PaLM, and Med-PaLM. Collectively, this broad spectrum of systems, which charts the evolution of MQA from initial pre-trained contextual encoders to today's billion-parameter, instruction-following architectures, offers a central empirical context against which the performance and results of our own model can be accurately and meaningfully interpreted.

As shown in Table V, the Mistral 7B-Instruct model achieves a diagnostic accuracy of 42.7%, outperforming all early transformer baselines including BERT-Base (34.3%), BioBERT-Base (34.1%), and RoBERTa-Large (35.0%), as well as the strongest pre-BERT variant, BioBERT-Large (36.7%). These results highlight that even without biomedical pretraining, a mid-scale, instruction-tuned model can offer stronger reasoning and factual comprehension than traditional transformer encoders trained on domain-specific quantities. When compared with more recent large-scale instruction-tuned models, such as PubMedGPT (50.3%), Flan-PaLM (67.6%), and Med-PaLM (up to 70.0%), the Mistral 7B-Instruct model understandably yields lower absolute accuracy. However, this performance gap primarily reflects the substantial differences in model scale, training data volume, and domain alignment rather than architectural limitations.

In contrast to these billion-parameter systems, the Mistral 7B-Instruct model operates within a lightweight, open-weight configuration, prioritising transparency, reproducibility, and fairness auditability criteria rarely satisfied by proprietary LLMs. Moreover, its stable demographic performance across the MedQA-Demog dataset underscores the model's potential as a controlled and ethically deployable foundation for future bias quantification research. These findings collectively demonstrate that while larger, domain-specific LLMs remain superior in raw diagnostic reasoning, open-weight instruction-tuned models like Mistral 7B-Instruct strike a meaningful balance between accuracy, interpretability, and fairness accountability, key priorities for responsible AI deployment in medical contexts.

TABLE VI
STATISTICAL COMPARISON OF DIAGNOSTIC ACCURACY ACROSS DEMOGRAPHIC GROUPS IN THE MEDQA-DEMOG DATASET.

Comparison	*Δ Accuracy*	*p (z-test)*	*Bonf. p*	*BH q*	*Cohen's h*	*Bootstrap 95 % CI*	*Bias Level*
Gender vs Original	- 0.025	0.219	0.657	0.293	- 0.050	[- 0.065, +0.015]	Low Bias
Race/Ethnicity vs Original	- 0.021	0.293	0.878	0.293	- 0.043	[- 0.060, +0.018]	Minimal Bias
Age vs Original	- 0.040	0.048	0.143	0.143	- 0.081	[- 0.080, - 0.000]	Low Bias

2) GROUP-SPECIFIC DISPARITIES AND SIGNIFICANCE TESTING

This section examines whether the demographic modifications introduced in MedQA-Demog, covering gender, race/ethnicity, and age, led to any statistically significant changes in diagnostic accuracy compared with the original USMLE questions. Each demographic subset included over 1,100 questions, providing sufficient statistical power (detectable difference ≈ 0.057 at 80 % power, $\alpha = 0.05$) to identify even small fairness-related effects. Across 4,659 evaluated items (1,272 original + 3,387 counterfactuals), the Mistral 7B-Instruct model demonstrated stable diagnostic reasoning under all demographic modifications. Table VI summarizes accuracy, Wilson 95 % CIs, and absolute differences versus baseline.

As shown in Table VI, the comparative statistical analysis across demographic groups confirms that the Mistral 7B-Instruct model maintains consistent diagnostic reasoning performance under all demographic perturbations. Accuracy differences between each demographic subset and the original MedQA items are minimal, with none exceeding a four-percentage-point deviation. The age-modified subset showed the largest observed reduction in accuracy ($\Delta = - 0.040$), yet this difference remains within the Low Bias range ($\Delta_{max} < 0.05$; see Table II). While the unadjusted z-test indicated a marginally significant result ($p = 0.048$), this effect became non-significant after applying Bonferroni ($p = 0.143$) and Benjamini-Hochberg (BH) corrections ($q = 0.143$), suggesting that the deviation was likely due to random variation rather than systematic demographic bias.

Similarly, the gender ($\Delta = - 0.025$) and race/ethnicity ($\Delta = - 0.021$) comparisons produced p-values well above conventional significance thresholds ($p = 0.219$ and $p = 0.293$, respectively), both classified as Low Bias and Minimal Bias. Their effect sizes were correspondingly small (Cohen's $h = - 0.050$ and $- 0.043$), and the bootstrap confidence intervals included zero, further confirming the absence of meaningful performance disparity.

Overall, across all comparisons, statistical corrections and confidence intervals consistently indicate no evidence of systematic demographic bias. The Mistral 7B-Instruct model's diagnostic accuracy appears robust and demographically invariant when evaluated on the MedQA-Demog benchmark.

3) MACRO PRECISION, RECALL, AND F1- SCORES

As shown in Table VII, macro-averaged precision, recall, and F_1-scores closely followed the overall accuracy trends observed earlier (see Section 3.1). Across all demographic variants, performance differences relative to the original MedQA subset remained below 0.05, confirming that the Mistral 7B-Instruct model preserved balanced predictive behaviour across gender, race/ethnicity, and age perturbations. Specifically, the original group achieved macro-precision of 0.426, macro-recall of 0.427, and macro-F1 of 0.426. The gender and race/ethnicity variants exhibited minor declines (-0.023 and -0.019 in F_1, respectively), whereas the age variant recorded a slightly larger decrease (–0.041), consistent with its modest drop in accuracy reported in Table III.

TABLE VII
MACRO-AVERAGED PRECISION, RECALL, AND F1 SCORES.

Group	*Macro-Precision*	*Macro-Recall*	*Macro-F1*
Original	0.426	0.427	0.426
Gender	0.403	0.399	0.400
Race/Ethnicity	0.407	0.407	0.405
Age	0.386	0.385	0.385

Importantly, these fluctuations remain within the Low Bias classification threshold ($\Delta_{max} < 0.05$; see Table II), suggesting that demographic changes apply minimal influence on the model's overall precision–recall balance.

Beyond aggregate scores, a closer examination of error structure revealed that the most frequent misclassifications involved confusion between semantically or clinically similar options. The top ten error transitions (truth → prediction) were dominated by minor label inversions such as B → A (298), C → A (272), C → D (253), and D → A (252). These symmetric confusion patterns suggest that the model's reasoning errors were non-systematic and semantically neutral, reflecting inherent uncertainty in complex medical reasoning rather than bias toward any particular demographic subgroup. Figure 4 visualizes the close alignment among macro-precision, recall, and F_1 scores across demographic variants, highlighting the model's consistent behaviour and fairness stability within the MedQA-Demog benchmark.

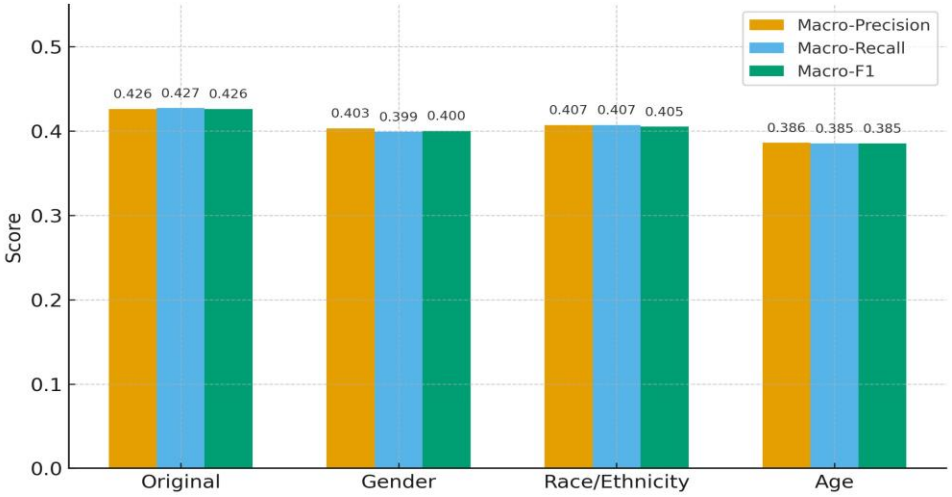

FIGURE 4. Macro-Precision, Macro-Recall, and Macro-F_1 scores across demographic groups (Original, Gender, Race/Ethnicity, and Age).

In summary, the macro-level evaluation supports earlier findings: performance stability and predictive symmetry are maintained across all demographic variations, highlighting the Mistral 7B-Instruct model's robustness and demographic fairness within the MedQA-Demog dataset.

Furthermore, to visualise diagnostic behaviour and assess whether demographic perturbations introduced systematic misclassification patterns, Figure 5 presents the aggregated confusion matrices for all evaluation subsets: the original MedQA questions and their gender-, race/ethnicity-, and age-modified counterfactuals. Across all panels, diagonal dominance is consistently maintained, indicating that the majority of predictions align with the correct diagnostic option. The off-diagonal cells representing errors show similar magnitudes and symmetrical dispersion, suggesting that most misclassifications stem from semantic proximity among clinically related options rather than from demographic bias. Importantly, no single group displays an inflated or depleted confusion region, reinforcing the earlier statistical finding that demographic edits did not introduce systematic reasoning drift in the model's responses.

A detailed examination of the confusion matrices in Figure 5 reveals that the model's misclassifications are largely confined to semantically adjacent diagnostic options rather than random or demographically driven errors. Across all groups, the ten most frequent error transitions follow consistent trends dominated by confusions between clinically probable alternatives such as B → A (298 cases), C → A (272), C → D (253), and D → A (252). These recurrent cross-predictions often involve conceptually overlapping diagnoses or therapeutically related conditions, suggesting that the model's uncertainty is epistemic (linked to medical reasoning ambiguity) rather than sociodemographic in nature. The relative symmetry of inverse transitions (e.g., A → B vs B → A) further supports this interpretation: no single answer category was disproportionately over- or under-predicted across any demographic subset. Collectively, this behaviour reinforces that observed variability arises from the intrinsic diagnostic complexity of the MedQA items rather than from sensitivity to gender, race/ethnicity, or age cues.

4) PREDICTION BALANCE AND OVER/UNDER-PREDICTION

To examine whether demographic perturbations induced any systematic preference for particular answer options, we analysed prediction balance, the relative deviation in predicted label frequency compared with the original MedQA distribution. This measure quantifies over-prediction (positive deviation) and under-prediction (negative deviation) for each of the four multiple-choice options (A–D). As shown in Table VIII, the Mistral 7B-Instruct model preserved a near-balanced prediction pattern across all demographic groups.

TABLE VIII
PREDICTION BALANCE AND OVER/UNDER-PREDICTION BY DEMOGRAPHIC GROUP.

Group	A	B	C	D
Original	+0.009	-0.019	-0.011	+0.021
Gender	+0.055	-0.028	-0.031	+0.004
Race/Ethnicity	+0.030	-0.019	-0.040	+0.029
Age	+0.027	-0.019	-0.025	+0.018

For the original dataset, deviations were minimal (ranging between - 0.019 and +0.021), indicating a well-distributed output probability across all answer choices. When demographic counterfactuals were introduced, slight fluctuations appeared but remained within expected random variation. The gender variant showed a modest

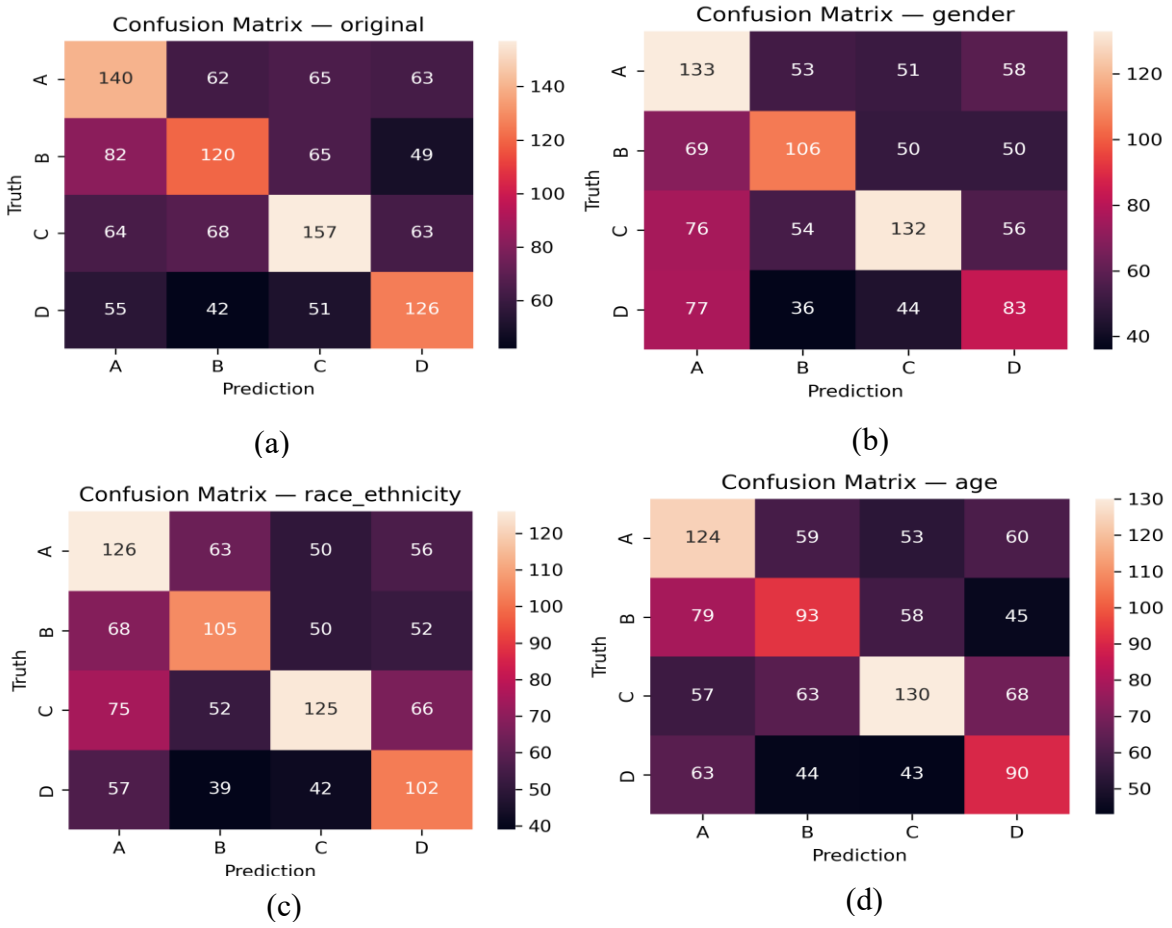

FIGURE 5. Confusion matrices for each demographic variant of the MedQA-Demog dataset: (a) Original, (b) Gender, (c) Race/Ethnicity, and (d) Age. Each matrix depicts predicted versus true answer distributions.

over-prediction for option A (+0.055) and mild under-prediction for C (- 0.031), while the race/ethnicity and age variants exhibited similarly small deviations (< ±0.04). No consistent directional bias (e.g., persistent preference for a specific label) was observed across groups, confirming that demographic edits did not distort the model's decision distribution. Overall, the distributions indicate stable response diversity across all counterfactual conditions. The absence of systematic over- or under-prediction patterns reinforces the earlier conclusion that the Mistral 7B-Instruct model maintained demographically neutral diagnostic reasoning, with fluctuations consistent with sampling noise rather than structural bias.

5) POWER AND MINIMUM DETECTABLE EFFECTS

To assess the statistical sensitivity of the fairness evaluation, we conducted a power analysis to estimate the smallest performance gap that could be detected with 80 % statistical power ($\alpha = 0.05$). This analysis ensures that any non-significant findings reported in previous sections are not due to insufficient sample size but rather reflect genuine performance parity across demographic variants. As shown in Table IX, each demographic comparison involved more than 1,100 counterfactual questions, matched against 1,272 original MedQA items. The Baseline Accuracy (*BaselineAcc*) represents the model's reference accuracy on the original dataset (0.427). The n_ref column denotes the number of baseline (original) samples, while *n_grp* refers to the sample size of each corresponding demographic variant (gender, race/ethnicity, or age). The final column, *MDE_abs_acc*, reports the minimum detectable effect, the smallest absolute accuracy difference that can be identified with 80 % power given the respective sample sizes.

Given these sample sizes, the minimum detectable absolute difference in accuracy was approximately 0.057 (5.7 percentage points) for all comparisons. As all observed disparities reported in Section 2 were smaller than this threshold (≤ 0.04), they fall below the detectable range, indicating that the apparent differences are statistically insignificant and unlikely to represent systematic demographic bias. In summary, this power analysis confirms that the non-significant fairness outcomes reported earlier reflect genuine model stability rather than limited data resolution. The Mistral 7B-Instruct model, therefore, demonstrates robust demographic neutrality within the MedQA-Demog evaluation framework.

TABLE IX
POWER ANALYSIS AND MINIMUM DETECTABLE ACCURACY GAPS (80 % POWER, A = 0.05).

Group	*BaselineAcc*	*n_ref*	*n_grp*	*MDE_abs_acc*
Gender	0.427	1272	1129	0.057
Race/Ethnicity	0.427	1272	1129	0.057
Age	0.427	1272	1129	0.057

IV. DISCUSSION

The results of this audit provide clear evidence that the Mistral 7B-Instruct model demonstrates robust demographic neutrality when applied to clinical question answering. Across 4,659 MedQA-Demog items, including systematic gender, race/ethnicity, and age perturbations, performance fluctuations remained minimal ($\Delta \leq 0.04$) and statistically non-significant. These findings indicate that open-weight LLMs can provide stable diagnostic reasoning despite controlled demographic variations, which is an encouraging result considering concerns regarding fairness in medical AI systems. A stratified robustness check by baseline question difficulty further confirms that this demographic stability is not driven by uniformly limited performance but persists across difficulty strata.

These findings should be interpreted in light of our design choice to treat demographic descriptors as fairness probes rather than clinically causal features, ensuring that observed effects reflect demographic framing sensitivity rather than medical risk modelling.

Compared with prior work, the present results show that demographic fairness is achievable even without large-scale parameter counts or domain-specific pre-training. Earlier transformer baselines such as BioBERT or RoBERTa-Large achieved below 40% accuracy on the MedQA benchmark [16], while instruction-tuned and domain-aligned models, such as Flan-PaLM and Med-PaLM reached 67-70 % accuracy with billions of parameters and extensive clinical alignment. Within this landscape, Mistral 7B-Instruct achieved 42.7%, outperforming early encoders despite being an open-source model with an order of magnitude fewer parameters. Notably, our findings complement prior work emphasising that fairness and reasoning quality are not strictly functions of model scale, but of controlled training and evaluation design [53]. The present study extends that insight by empirically demonstrating demographic robustness within a fully reproducible, open-weight framework.

Error-structure analysis further supports this interpretation. Confusion matrices and prediction-balance measures revealed symmetrical misclassifications dominated by semantically adjacent diagnostic options rather than demographically patterned errors. Frequent transitions such as B → A or C → D reflected epistemic uncertainty within the clinical content, consistent across all demographic subsets. The absence of systematic over- or under-prediction trends (Table VIII) indicates that the model's diagnostic behaviour is shaped primarily by inherent case difficulty rather than demographic perturbation. In practical terms, this means that apparent performance differences between demographic groups are statistically indistinguishable from random sampling noise.

From a methodological standpoint, the results validate the proposed MedQA-Demog audit framework as a transparent, replicable approach for bias evaluation in medical LLMs. The framework's integration of counterfactual augmentation, Wilson confidence intervals, and effect-size reporting (Cohen's h) ensures interpretability and quantitative accuracy, aligning with emerging standards for fairness auditing. Importantly, the power analysis (Table IX) confirmed that all observed gaps were below the minimum detectable threshold (≈ 0.057), indicating that non-significance reflects genuine model stability rather than insufficient data.

These outcomes have several broader implications. First, they highlight the potential of open-weight models such as Mistral to serve as transparent research baselines for fairness benchmarking, enabling reproducibility and external verification absent in proprietary systems. Second, they show that bias auditing frameworks can and should extend beyond aggregate accuracy to include structured statistical testing, effect-size interpretation, and visual inspection of misclassification patterns. Third, the minimal demographic sensitivity observed here suggests that instruction-tuned open models may already possess sufficient contextual grounding to generalise equitably across basic patient characteristics, though further evaluation on free-text clinical notes and multi-modal inputs is necessary. Finally, while Mistral 7B-Instruct's overall diagnostic accuracy remains lower than domain-aligned giants like Med-PaLM 2 or GPT-4, its transparency, efficiency, and fairness stability position it as a viable foundation for academic and clinical research.

The primary contribution of this work lies in methodological rigor for fairness auditing rather than optimisation of diagnostic accuracy, which remains an important direction for future studies applying the proposed framework to higher-performing medical LLMs. While this study focuses on diagnostic decision consistency under demographic counterfactuals, future extensions of the framework could incorporate analyses of explanation tone, certainty calibration, and recommendation strength to assess communicative fairness in patient or clinician-facing settings.

In the present study, demographic attributes are evaluated independently in order to preserve label invariance and

avoid introducing clinically implausible or confounded combinations within the MedQA vignettes; this design choice ensures that observed effects can be attributed to demographic framing rather than unintended changes in clinical semantics. Future work should expand this auditing framework to include intersectional attributes (e.g., age × gender), reasoning-trace analysis, and reinforcement-based fairness tuning to ensure equitable, trustworthy clinical deployment of medical LLMs. Moreover, the present study intentionally employs a fixed, deterministic prompt template to isolate demographic effects and ensure full reproducibility of the fairness audit. This design choice enables controlled comparison across demographic counterfactuals without introducing additional variability from prompt engineering or external knowledge retrieval. Future extensions of this framework could evaluate demographic robustness under alternative prompt templates and simple retrieval-augmented configurations to better approximate real-world clinical decision-support deployments.

In summary, the present findings confirm that demographic perturbations do not meaningfully change the diagnostic reasoning of the Mistral 7B-Instruct model within the MedQA-Demog framework. The proposed audit framework provides a robust and transparent foundation for fairness evaluation in clinical LLMs and demonstrates that reproducible, open-source infrastructures can achieve both methodological rigour and ethical accountability. These outcomes directly inform the next section, which outlines the broader implications, limitations, and future pathways toward trustworthy, bias-aware medical AI systems.

V. CONCLUSION

This work presents one of the first systematic evaluations of demographic fairness in open-weight medical language models. Through the development of MedQA-Demog, a label-invariant, counterfactually augmented version of the MedQA-USMLE dataset and its deployment within a fully local, transparent inference pipeline, we provide strong evidence that the Mistral 7B-Instruct model shows stable diagnostic reasoning across patient gender, race/ethnicity, and age. Accuracy fluctuations across all demographic variants remained within ± 0.04 and were statistically non-significant (all $p > 0.05$). Effect sizes were minimal ($|h| < 0.1$), and bootstrap confidence intervals consistently included zero, confirming that residual variation reflected stochastic rather than systematic bias. These results demonstrate that open-source instruction-tuned LLMs can achieve robust demo-graphic neutrality when evaluated under controlled and reproducible conditions.

Beyond model performance, this study establishes a transparent methodological framework for bias auditing, incorporating structured counterfactual augmentation, deterministic inference, power analysis, and quantitative bias categorisation. Together, these elements form a reproducible, extensible standard for assessing fairness in clinical AI systems. Future research should extend this framework to multimodal medical tasks, intersectional demographic attributes (e.g., age × gender), and longitudinal fine-tuning protocols designed to reinforce fairness while preserving diagnostic validity. By embedding such reproducible fairness auditing practices in the model-development lifecycle, we take a critical step toward trustworthy, equitable, and accountable deployment of medical LLMs.

REFERENCES

[1] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, arXiv:2108.07258.
[2] J. Clusmann, F.R. Kolbinger, H.S. Muti *et al.* "The future landscape of large language models in medicine." *Commun Med* **3**, 141 (2023). https://doi.org/10.1038/s43856-023-00370-1
[3] J. Achiam et al., "GPT-4 technical report," 2023, arXiv:2303.08774
[4] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," Nature, vol. 616, no. 7956, pp. 259–265, Apr. 2023
[5] H. Ali, J. Qadir, Z. Shah, T. Alam, and M. Househ, ''ChatGPT and large language models (LLMs) in healthcare: Opportunities and risks,'' TechRxiv, vol. 1, pp. 1–14, Jul. 2023.
[6] F. S.P.A.A. and D. Wickramaarachchi, "Large Language Model (LLM) Support for Preliminary Consultation in Healthcare," in *2025 5th International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka, 2025, pp. 1–6, doi: 10.1109/ICARC64760.2025.10963287.
[7] M. Benary, XD. Wang, M. Schmidt, et al. "Leveraging Large Language Models for Decision Support in Personalized Oncology." *JAMA Network Open.* 2023;6(11): e2343689. doi:10.1001/jamanetworkopen.2023.43689
[8] D. V. Veen, C. V. Uden, and L. Blankemeier, ''Clinical text summarization: Adapting large language models can outperform human experts,'' Research Square, vol. 1, pp. 1–26, Sep. 2023
[9] D. D., T. D. Rajkumar, and D. Balakrishnan, "Optimized Medical Recommendation System Utilizing Large Language Models for Enhanced Question Answering Performance," in *2024 Global Conference on Communications and Information Technologies (GCCIT)*, Bangalore, India, 2024, pp. 1–5, doi: 10.1109/GCCIT63234.2024.10862477.
[10] Ji Y, Zhang H, Wang Y. "Bias evaluation and mitigation in retrieval-augmented medical question-answering systems." arXiv preprint arXiv:2503.15454. 2025 Mar 19.
[11] R. Poulain, H. Fayyaz, R. Beheshti. "Bias patterns in the application of LLMs for clinical decision support: A comprehensive study." arXiv preprint arXiv:2404.15149. 2024 Apr 23.
[12] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," PLOS Digit. Health, vol. 2, no. 2, Feb. 2023, Art. no. e0000198.
[13] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, and D. Chartash, "How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment," JMIR Med. Educ., vol. 9, Feb. 2023, Art. no. e45312.
[14] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," 2023, arXiv:2303.13375
[15] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, and D. Neal, "Towards expertlevel

medical question answering with large language models," 2023, arXiv:2305.09617
[16] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? A large-scale open domain question answering dataset from medical exams," Appl. Sci., vol. 11, no. 14, p. 6421, Jul. 2021.
[17] Q. Peng, J. Liu, Q. Zou, X. Chen, Z. Zhong, Z. Wang, J. Xie, Y. Cai, and Q. Li, "Integration of multi-source medical data for medical diagnosis question answering," IEEE Transactions on Medical Imaging, 2024.
[18] K. Shang, C.-H. Chang, and C. C. Yang, "Collaboration among multiple large language models for medical question answering," *arXiv* preprint arXiv:2505.16648, 2025. [Online]. Available: https://arxiv.org/abs/2505.16648. Accessed: Nov. 3, 2025.
[19] A. Ghosh and K. Deepa, "QueryMintAI: Multipurpose Multimodal Large Language Models for Personal Data," IEEE Access, vol. 12, pp. 144631-144651, 2024.
[20] I. O. Gallegos, RA. Rossi, J. Barrow, MM. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, NK. Ahmed, "Bias and fairness in large language models: A survey," 2023, arXiv:2309.00770
[21] Y. Xiao, J. Huang, R. He, J. Xiao, MR. Mousavi, Y. Liu, K. Li, Z. Chen, JM. Zhang. "AMQA: An Adversarial Dataset for Benchmarking Bias of LLMs in Medicine and Healthcare." arXiv preprint arXiv:2505.19562. 2025 May 26.
[22] B. Ni, Z. Liu, L. Wang, et al., "Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2502.06872, 2025
[23] P. Spitzer, D. Hendriks, J. Rudolph, S. Schlaeger, J. Ricke, N. Kühl, BF. Hoppe, S. Feuerriegel. "The effect of medical explanations from large language models on diagnostic decisions in radiology." medRxiv. 2025 Mar 6:2025-03.
[24] M. Raj, V.A. Ravindran, "A Comparative Study on Generative Artificial Intelligence by Evaluating Multiple Large Language Models for Guidance to Parents Toward Pediatric Dentistry: A Multimodal Comparative LLM Study." Journal of International Oral Health.:10-4103.
[25] A. Warrier, R. Singh, A. Haleem, H. Zaki, JA. Eloy. "The comparative diagnostic capability of large language models in otolaryngology." The Laryngoscope, Wiley, 2024 Sep;134(9):3997-4002.
[26] D. Pawar and S. Phansalkar, "MindWellQA: A semantically enriched evidence-based QA system for psychological disorders," *IEEE Access*, early access, 2025, doi: 10.1109/ACCESS.2025.3627880.
[27] J. A. Omiye, J. C. Lester, S. Spichak, V. Rotemberg, and R. Daneshjou, "Large language models propagate race-based medicine," NPJ Digit. Med., vol. 6, no. 1, 2023, Art. no. 195.
[28] T. Zack, E. Lehman, M. Suzgun, J. A. Rodriguez, L. A. Celi, J. Gichoya, D. Jurafsky, P. Szolovits, D. W. Bates, R.-E. E. Abdulnour et al., "Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study," The Lancet Digital Health, vol. 6, no. 1, pp. e12–e22, 2024.
[29] SR. Pfohl, H. Cole-Lewis, R. Sayres, D. Neal, M. Asiedu, A. Dieng, N. Tomasev, QM. Rashid, S. Azizi, N. Rostamzadeh, LG. McCoy. "A toolbox for surfacing health equity harms and biases in large language models." Nature Medicine. 2024 Dec;30(12):3590-600
[30] MD. Abràmoff, ME. Tarver, N. Loyo-Berrios, S. Trujillo, D. Char, Obermeyer Z, Eydelman MB, Foundational Principles of Ophthalmic Imaging and Algorithmic Interpretation Working Group of the Collaborative Community for Ophthalmic Imaging Foundation, Washington, DC, Maisel WH. Considerations for addressing bias in artificial intelligence for health equity. NPJ digital medicine. 2023 Sep 12;6(1):170.
[31] M. Mittermaier, M. M. Raza, and J. C. Kvedar, "Bias in AI-based models for medical applications: Challenges and mitigation strategies," NPJ Digit. Med., vol. 6, no. 1, Jun. 2023, doi: 10.1109/ACCESS.2023.10601195
[32] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, ''PubMedQA: A dataset for biomedical research question answering,'' 2019, arXiv:1909.06146.
[33] K. Benkirane, J. Kay, M. Perez-Ortiz. "How Can We Diagnose and Treat Bias in Large Language Models for Clinical Decision-Making?." arXiv preprint arXiv:2410.16574. 2024 Oct 21.
[34] A. Q. Jiang et al, "Mistral 7B", ArXiv e-prints, 2023. doi:10.48550/arXiv.2310.06825.
[35] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, arXiv:2307.09288.
[36] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Mach. Intell., vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.
[37] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in Proc. ACM Conf. Fairness, Accountability, Transparency, 2021, pp. 610–623.
[38] M. Alkhalaf et al., "Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records," *Journal of Biomedical Informatics*, p. 104662, 2024.
[39] L. Riedemann, M. Labonne, and S. Gilbert, "The path forward for large language models in medicine is open," *npj Digital Medicine*, vol. 7, Art. no. 339, 2024, doi: 10.1038/s41746-024-01344-w.
[40] I. Y. Chen, S. Joshi, and M. Ghassemi, "Treating health disparities with artificial intelligence," Nature Med., vol. 26, no. 1, pp. 16–17, Jan. 2020.
[41] A. d'Elia, M, Gabbay, S, Rodgers, C. Kierans, E. Jones, I. Durrani, A. Thomas, L. Frith. "Artificial intelligence and health inequities in primary care: a systematic scoping review and framework." Family Medicine and Community Health. 2022 Nov 30;10(Suppl 1):e001670.
[42] Mistral-7B-Instruct-v0.3. Available online: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3 (accessed on 11 October 2025).
[43] Ollama API Documentation. *Available online:* https://docs.ollama.com/api *(accessed on 11 October 2025).*
[44] J. Cohen, Statistical Power Analysis for the Behavioral Sciences. Nashville, TN, USA: Abingdon, 1988.
[45] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi, and I. Golani, ''Controlling the false discovery rate in behavior genetics research,'' Behavioural Brain Res., vol. 125, nos. 1–2, pp. 279–284, Nov. 2001. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0166432801002972
[46] H. O. Ahmed and A. K. Nandi, ''High performance breast cancer diagnosis from mammograms using mixture of experts with EfficientNet features (MoEffNet),'' IEEE Access, vol. 12, pp. 133703–133725, 2024, doi: 10.1109/ACCESS.2024.3461360.
[47] E. B. Wilson, ''Probable inference, the law of succession, and statistical inference,'' J. Amer. Stat. Assoc., vol. 22, no. 158, pp. 209–212, Jun. 1927.
[48] D. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, 2nd ed. London, U.K.: Chapman & Hall/CRC, 2006.
[49] J. L. Fleiss, B. Levin, and M. C. Paik. Statistical methods for rates and proportions. John Wiley & Sons, Inc., New Jersey, 2003. doi: 10. 1002/0471445428
[50] L. Seyyed-Kalantari et al., "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," Nature Med., vol. 27, no. 12, pp. 2176–2182, 2021.
[51] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in Proc. AAAI/ACM Conf. AI, Ethics, Soc., 2018, pp. 335–340.
[52] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, ''A survey on bias and fairness in machine learning,'' ACM Comput. Surv., vol. 54, no. 6, pp. 1–35, Jul. 2022.
[53] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, no. 7972, pp. 172–180, 2023.

HOSAMELDIN O. A. AHMED received a PhD degree in electronic and computer engineering from Brunel University London. He is currently a Distinguished Researcher at The Sir Peter Rigby Digital Futures Institute, Aston University, U.K. His interdisciplinary research is at the intersection of artificial intelligence (AI), multimodal medical imaging, and extended reality (XR). Dr Ahmed has pioneered the development of AI and deep learning frameworks for the early detection of lung and breast cancer, with a specialized focus on applying deep transfer learning to mammography, CT scan, X-ray, and Ultrasound images.

He has led projects integrating AI with electronic health records for NHS-interoperable clinical decision support and is a recognized innovator in XR, having designed immersive applications for medical training and real-time physiological data visualization. His research portfolio extends significantly to Industry 4.0, where he has developed machine-learning-driven solutions for predictive maintenance and condition monitoring of rotating machinery, as detailed in his co-authored book, *Condition Monitoring with Vibration Signals* (IEEE-Wiley, 2020). His scholarly work also encompasses internet addiction disorder detection and digital heritage preservation through image inpainting and 3D visual interaction. His current research interests include generative AI for synthetic data, explainable AI (XAI) in healthcare, and the development of AI-powered multimodal systems for next-generation digital health.

Professor Abdul Hamid Sadka. is a distinguished academic and researcher specializing in artificial intelligence (AI), visual computing, and digital innovation. He joined Aston University in 2023 as the founding Director of The Sir Peter Rigby Digital Futures Institute, after an extensive career at Brunel University London, where he served as Director of the Institute of Digital Futures, founding Director of the Brunel Digital Science and Technology Hub, Director of the Centre for Media Communications Researcher, and Head of Electronic, Computer Engineering and Digital Media. With nearly 30 years of academic and research leadership and cross-sector collaborations, Professor Sadka has made significant contributions to AI-enabled visual media technologies and intelligent multimodal systems. His expertise spans machine learning, image and video processing, immersive and autonomous systems, and digital health applications. He has successfully led large-scale interdisciplinary projects addressing healthcare, engineering, and security challenges, attracting over £20 million in competitive research funding. Professor Sadka is also the author of the Wiley book Compressed Video Communications (Wiley Online Library), a seminal work on video compression and transmission technologies that continues to inform research and industry practices. He has supervised more than 50 PhD and postdoctoral researchers to completion and published extensively in high-impact journals.

He has an extensive record of collaborations with industry and public-sector partners, developing AI-driven, multimodal, and immersive solutions that bridge fundamental research with real-world innovation across healthcare, manufacturing, digital media, and professional services.