DUCX: Decomposing Unfairness in Tool-Using Chest X-ray Agents

Zikang Xu[1*], Ruinan Jin[23*], and Xiaoxiao Li[23]

[1] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Anhui, China
[2] The University of British Columbia, Vancouver, BC V6Z 1Z4, Canada
[3] Vector Institute, Toronto, ON M5G 1M1, Canada
zikangxu@mail.ustc.edu.cn, ruinanjin@alumni.ubc.ca, xiaoxiao.li@ece.ubc.ca

Abstract. Tool-using medical agents can improve chest X-ray question answering by orchestrating specialized vision and language modules, but this added pipeline complexity also creates new pathways for demographic bias beyond standalone models. We present DUCX (**D**ecomposing **U**nfairness in **C**hest **X**-ray agents), a systematic audit of chest X-ray agents instantiated with MedRAX. To localize where disparities arise, we introduce a stage-wise fairness decomposition that separates *end-to-end bias* from three agent-specific sources: *tool exposure bias* (utility gaps conditioned on tool presence), *tool transition bias* (subgroup differences in tool-routing patterns), and *model reasoning bias* (subgroup differences in synthesis behaviors). Extensive experiments on tool-used based agentic frameworks across five driver backbones reveal that (i) demographic gaps persist in end-to-end performance, with equalized odds up to 20.79%, and the lowest fairness-utility tradeoff down to 28.65%, and (ii) intermediate behaviors, tool usage, transition patterns, and reasoning traces exhibit distinct subgroup disparities that are not predictable from end-to-end evaluation alone (e.g., conditioned on segmentation-tool availability, the subgroup utility gap reaches as high as 50%). Our findings underscore the need for process-level fairness auditing and debiasing to ensure the equitable deployment of clinical agentic systems. Code is available here.

Keywords: Fairness in Medical Agents · Fairness in Agents · Medical AI Agents

1 Introduction

Artificial intelligence (AI) is increasingly integrated into medical imaging, supporting tasks from disease detection to clinical decision assistance [32]. Despite

*Ruinan Jin and Zikang Xu contributed equally to this paper.
Corresponding author: Xiaoxiao Li

strong performance gains from vision and vision-language models, growing evidence shows that medical AI can exhibit demographic disparities across patient subgroups (e.g., age and gender), raising concerns about safety, reliability, and equitable deployment [26,20].

Most fairness studies in medical AI have focused on *standalone* models [26,12], including convolutional networks [27], vision foundation models (FMs) [12], and large multimodal models (LMMs) [25,13]. In parallel, medical AI systems are rapidly shifting toward *agentic* architectures that solve tasks via *multi-step pipelines*, orchestrating multiple tools and models under the control of a large language model (LLM) planner [5,24,23]. By dynamically selecting tools such as classifiers, segmentators, report generators, and visual question answering (VQA) modules, agents can improve flexibility and interpretability, but they also introduce new pathways through which unfairness may arise and propagate.

In fact, *fairness in agentic systems is not a direct extension of model-level fairness*. Standalone fairness audits treat a model as a single decision function and measure disparities in final predictions. In contrast, agentic systems spend more effort on how to reason user queries, planning actions, and communicating among different models, resulting in *new fairness failure modes* that do not exist in single-step models.

However, current fairness evaluations rarely audit the agent's internal process. Without process-level attribution, it is hard to diagnose whether disparities are *inherited* from particular vision-language tools, *introduced* by the planner's tool-selection and transition behavior, or *amplified* during final response generation. We address this gap by conducting a *stage-wise fairness decomposition* of tool-using chest X-ray agents for multiple-choice question answering.

We study MedRAX [5], because it offers (i) *complete trajectory observability* for auditing and decomposing, and (ii) a *controlled single-agent pipeline* where decomposition into *tool-exposure*, *tool-transition*, and *reasoning* disparities is well-defined, and intervention targets are modular. Using CheXAgent-Bench and our curated MIMIC-FairnessVQA dataset, we evaluate fairness both at the *outcome level* and at the *process level*, decomposing agent unfairness into (i) *tool-exposure bias*, (ii) *tool-transition bias*, and (iii) *LLM reasoning bias*. This decomposition yields a straightforward map from observed disparity to the agent stage most responsible, directly motivating our evaluation framework in Sec. 2.3.

Our contributions are threefold. (1) We perform the first systematic demographic fairness evaluation of MedRAX-style chest X-ray agents across five driver LLMs under a unified setup. (2) We propose `DUCX` (**D**ecomposing **U**nfairness in **C**hest **X**-ray agents), a stage-wise framework with metrics that attribute disparities to *tool exposure*, *tool transition*, and *LLM reasoning*. (3) We curate *MIMIC-FairnessVQA*, a demographic-aware benchmark with standardized (image, multi-choice question, demographics) instances for chest X-ray agents.

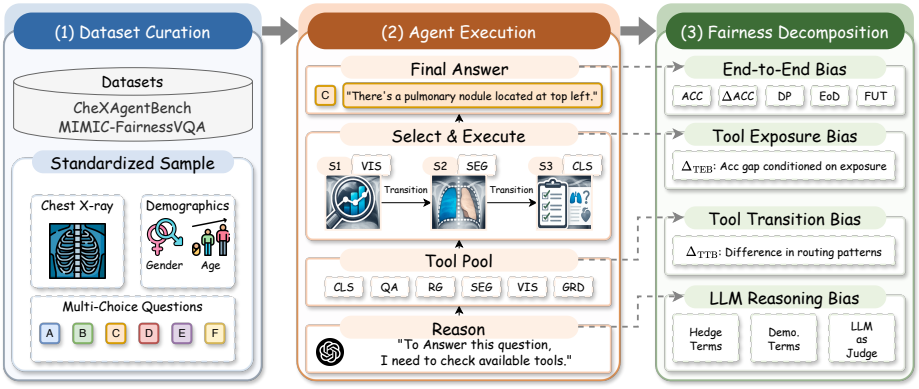

Fig. 1. Overview of DUCX. (1) Dataset curation: we use CheXAgentBench and curate MIMIC-FairnessVQA into a standardized form. (2) MedRAX agent execution (ReAct): a driver LLM iteratively reasons, selects tools, and synthesizes a final answer, where we highlight multiple points where bias can be introduced (tool exposure, tool transitions, and final synthesis). (3) Fairness decomposition: We evaluate *end-to-end fairness* (ACC, ΔACC, DP, EoD, FUT) and decompose it into *tool exposure bias* (gaps conditioned on tool presence), *tool transition bias* (gaps in tool routing), and *LLM reasoning bias* (gaps in synthesis behaviors).

2 DUCX

2.1 Preliminaries

Agentic systems in medical imaging. Traditional *standalone* medical imaging models map an input image directly to an *end-to-end output* (e.g., a label or report), so fairness evaluation typically concerns only the final prediction [26,14]. In contrast, recent systems are increasingly *agentic* [5,22,17]: an LLM planner decomposes a query into multiple steps and autonomously invokes specialized tools (e.g., classifiers, segmenters, report generators, grounding, visualization, or VQA modules) to produce the final answer, without human intervention as shown in Fig. 1. For example, MedRAX [5] is a representative tool-using X-ray agent framework that follows a ReAct-style loop [29]: the agent iteratively *reasons* about the next step, *acts* by calling a tool or generating an intermediate response, and *observes* the result to update its context until completion. This structured execution trace *exposes multiple decision points beyond the final output*, enabling the motivation for stage-wise analysis and decomposition of unfairness in agentic systems.

Fairness in medical imaging. Prior work on fairness in medical imaging has largely evaluated *standalone* predictors, operationalizing equity as comparable performance and error behavior across demographic groups [30,11,21,18,7]. *Group fairness* measures each subgroup's *end-to-end performance* (e.g., accuracy) and summarizes disparity by the *gap* between groups [10]; These notions are well-suited to single-step models, but emerging *agentic* pipelines produce an-

swers via intermediate decisions (e.g., tool exposure, tool transition, and LLM synthesis), creating additional pathways for disparities that are not visible from final outcomes alone. This motivates auditing fairness not only at the outcome level, but also at the *process level*: whether different groups induce systematically different tool-routing policies or generation behaviors that may ultimately drive end-to-end gaps.

2.2 Problem Setting and Agent Execution

Let $\mathcal{D} = \{(x_i, q_i, y_i, a_i)\}_{i=1}^{N}$ denote a dataset of chest X-ray questions, where x is an image (and optional clinical context), q is a question with K candidate answers, $y \in \{1, \ldots, K\}$ is the ground-truth option index, and a is a sensitive attribute (e.g., gender). Let $g \in \mathcal{G}$ index sensitive subgroups induced by a (e.g., $\mathcal{G} = \{\text{Female}, \text{Male}\}$ or $\{\text{Age} < 60, \text{Age} \geq 60\}$). An agent produces a final prediction $\hat{y}$ by executing a tool-augmented trajectory τ:

$$\tau = \big(s_0, (t_1, o_1), \ldots, (t_T, o_T), \hat{y}\big), \tag{1}$$

where s_0 is the initial state formed from (x, q), $t_j \in \mathcal{T}$ denotes the selected tool at step j from a tool set $\mathcal{T}$, o_j is the corresponding tool output (observation), and T is the number of tool calls before producing the final answer. We assume access to execution logs containing (t_j, o_j) and the final natural-language response.

Driver LLMs. Five LLMs including LLaMA3.1-8B [8], Ministral-3-8B [19], Qwen3VL-8B [2], Qwen3-8B [28], and Gemini3-Flash [3], are used to drive the agent, to evaluate potential unfairness introduced by the driver LLM.

Tools Pool. Six categories of tools are included in the tools pool for calling, including classifier (CLS), visual question answering module (QA), report generator (RG), segmentator (SEG), visualizer (VIS), and phrase grounding module (GRD).

2.3 Fairness Decomposition

End-to-end bias. Three types of metrics are adopted following [14], i.e., (i) *Utility measurement*: accuracy (ACC); (ii) *Fairness measurement*: delta accuracy (ΔACC), demographic parity (DP), Equalized Odds (EoD) [10]; (iii) *Trade-off measurement*: fairness-utility trade-off score (FUT) [14].

Tool-exposure bias. *Motivation:* even when the agent uses the *same* tool, that tool may deliver *unequal downstream utility across groups* (e.g., due to training imbalance), which then propagates to the final decision. *Definition:* for a tool A, let $E_A(\tau) = \mathbb{1}[A \in \tau]$ indicate whether A is used at least once. We define tool-exposure bias as the *subgroup accuracy gap conditioned on exposure*:

$$\Delta_{\text{TEB}}(A) = \text{Acc}(g{=}g_1 \mid E_A(\tau){=}1) - \text{Acc}(g{=}g_2 \mid E_A(\tau){=}1)\,. \tag{2}$$

Evaluation: we filter to instances with $E_A(\tau) = 1$, compute subgroup accuracies, and report $\Delta_{\text{TEB}}(A)$ (and $|\Delta_{\text{TEB}}(A)|$). *To avoid confounding with routing*, we also report each group's *exposure rate* $P(E_A(\tau) = 1 \mid g)$.

Tool-transition bias. *Motivation:* unfairness can arise even with identical tools if the LLM planner *routes different demographics through systematically different tool chains* (e.g., longer or less reliable sequences example in the introduction). *Definition:* let the tools pool be $\mathcal{T} = \{A_1, \ldots, A_T\}$. For each subgroup g, we estimate a *Markov transition matrix* $P^{(g)} \in \mathbb{R}^{T \times T}$:

$$p_{i,j}^{(g)} = P(t_{k+1} = A_j \mid t_k = A_i, g), \qquad \sum_{j=1}^{T} p_{i,j}^{(g)} = 1. \tag{3}$$

We define tool-transition bias as the *difference in routing patterns*:

$$\Delta_{\mathrm{TTB}} = P^{(g_1)} - P^{(g_2)}. \tag{4}$$

Evaluation: entries of Δ_{TTB} identify which tool-to-tool transitions are *more frequent for one group*, highlighting planning-level disparity beyond final accuracy.

LLM reasoning bias. *Motivation:* even with identical trajectories, unfairness may manifest even when tools and answers are comparable, because the driver LLM can produce *demographic-dependent reasoning quality and communication style* (e.g., different uncertainty expression or demographic framing). We therefore decompose *synthesis-level disparities* directly from the agent's final response text. *Definition:* For a sensitive attribute with groups $\mathcal{G}$ and a response-level feature f, we compute subgroup means $\mu_g = \mathbb{E}[f \mid g]$ and define the reasoning-bias gap as:

$$\Delta f = \max_{g \in \mathcal{G}} \mu_g - \min_{g \in \mathcal{G}} \mu_g. \tag{5}$$

Evaluation: We report three complementary gaps: (i) *JudgeGap*: ΔJudgeGap, where f is a *reasoning quality score* produced by an external LLM judge given the *question* and the agent's final response. We adapt the exact prompt in [31] for the LLM judge. (ii) *Hedge* [1,16]: ΔHedge, where f counts *hedging cues* (e.g., *may, might, possibly, likely, appears*) via case-insensitive pattern matching; (iii) *Demographic* [4,9]: ΔDemo., where f counts *explicit demographic terms* (e.g., *male, female, elderly, young*) via pattern matching. Above, JudgeGap captures overall reasoning quality disparity, while ΔHedge and ΔDemo. capture systematic differences in uncertainty expression and demographic framing.

2.4 Dataset Curation

Evaluating fairness in *agentic* chest X-ray systems requires *agent-suitable* questions that support multi-step tool use, include sensitive attributes, and allow automatic scoring. Such datasets are scarce; to our knowledge, only ChestAgentBench [5] provides both agent-oriented questions and patient demographics. We therefore use ChestAgentBench and curate a new MIMIC-based benchmark.

CheXAgentBench [5]. It contains 2,500 diagnosis questions from 675 expert-curated clinical cases (Eurorad[1]). Each sample includes a chest X-ray, patient

[1] https://www.eurorad.org/

demographics, and a six-choice question. The dataset is available via HuggingFace [6].

MIMIC-FairnessVQA[2]**.** We curate MIMIC-FairnessVQA from MIMIC-CXR [15] to enable demographic-aware evaluation at scale: (i) *Demographics-balanced sampling*: we sample 400 studies balanced by gender and age (threshold 60); (ii) *Information extraction*: we extract metadata, the final report, and the X-ray image; (iii) *LLM-based question generation*: we prompt LLM (using a MedRAX-style template) to generate six-choice questions with explanations across five task focuses per study, yielding 2,000 instances. We manually verify answer correctness.

3 Results

In this section, we systematically inspect unfairness in agent instruction following the scheme in Sec. 2.3, including end-to-end bias, tool-exposure bias, tool-transition bias, and LLM-reasoning bias.

Implementation details. We follow the open-source MedRAX codebase for evaluation and use DeepSeek as the judge model. We assess statistical significance with non-parametric bootstrapping (1,000 resamples) and report confidence intervals (CIs). For sensitive attributes, we consider gender (male vs. female) and age (young <60 vs. old ≥60).

Table 1. Overview of the end-to-end fairness performance. Mean_{CI} is reported. All values are in percentage. Best and Second Best in each group are highlighted.

Configuration		CheXAgentBench					MIMIC-FairnessVQA				
Attr.	LLM	ACC ↑	ΔACC ↓	DP ↓	EoD↓	FUT ↑	ACC ↑	ΔACC ↓	DP ↓	EoD↓	FUT ↑
Gender	**LLaMA3.1**	$50.83_{[48.88,52.76]}$	$1.53_{[0.06,4.45]}$	$5.04_{[2.32,7.93]}$	$15.10_{[5.63,26.98]}$	$50.29_{[48.20,52.30]}$	$32.89_{[30.85,34.95]}$	$2.00_{[0.10,5.86]}$	$2.99_{[1.12,6.28]}$	$11.45_{[5.83,18.97]}$	$32.43_{[30.39,34.56]}$
	Ministral-3	$51.48_{[49.72,53.36]}$	$1.68_{[0.06,4.40]}$	$3.66_{[1.48,6.73]}$	$10.48_{[4.12,19.60]}$	$50.85_{[48.75,52.88]}$	$29.42_{[27.35,31.50]}$	$2.30_{[0.11,6.07]}$	$3.35_{[1.39,5.86]}$	$11.21_{[5.39,18.36]}$	$28.95_{[26.86,31.07]}$
	Qwen3VL	$67.26_{[65.11,69.28]}$	$3.21_{[0.21,7.11]}$	$4.47_{[1.42,8.12]}$	$11.28_{[4.87,22.64]}$	$65.98_{[63.41,68.45]}$	$54.75_{[52.07,57.39]}$	$2.17_{[0.11,6.29]}$	$4.22_{[1.74,7.56]}$	$11.47_{[4.93,20.61]}$	$53.92_{[51.02,56.77]}$
	Qwen3	$69.10_{[67.24,70.92]}$	$1.92_{[0.09,5.07]}$	$3.72_{[1.52,6.62]}$	$9.16_{[3.45,19.62]}$	$68.09_{[65.74,70.32]}$	$49.89_{[47.55,52.10]}$	$1.90_{[0.09,5.30]}$	$3.11_{[1.14,5.98]}$	$11.32_{[5.06,19.56]}$	$49.23_{[46.77,51.37]}$
	Gemini3	$50.60_{[48.64,52.64]}$	$3.62_{[0.36,7.47]}$	$4.99_{[2.61,7.99]}$	$9.56_{[4.83,16.61]}$	$49.15_{[46.82,51.59]}$	$49.69_{[47.50,51.95]}$	$1.99_{[0.08,5.31]}$	$3.13_{[1.36,5.38]}$	$11.01_{[5.31,19.10]}$	$49.01_{[46.54,51.48]}$
Age	**LLaMA3.1**	$50.83_{[48.88,52.76]}$	$2.66_{[0.13,6.45]}$	$2.92_{[1.00,5.69]}$	$13.58_{[5.91,23.11]}$	$49.47_{[46.60,51.87]}$	$32.89_{[30.85,34.95]}$	$2.27_{[0.08,6.00]}$	$2.82_{[1.22,5.76]}$	$9.63_{[4.48,16.65]}$	$32.40_{[30.30,34.53]}$
	Ministral-3	$51.48_{[49.72,53.36]}$	$1.89_{[0.07,5.33]}$	$3.00_{[1.16,5.69]}$	$12.08_{[4.85,21.98]}$	$50.60_{[48.02,52.81]}$	$29.42_{[27.35,31.50]}$	$3.55_{[0.22,7.50]}$	$4.73_{[2.32,7.70]}$	$11.38_{[5.62,18.52]}$	$28.65_{[26.53,30.74]}$
	Qwen3VL	$67.26_{[65.11,69.28]}$	$2.10_{[0.06,5.99]}$	$4.27_{[1.74,7.52]}$	$20.79_{[8.91,33.72]}$	$66.14_{[62.90,68.65]}$	$54.75_{[52.07,57.39]}$	$2.18_{[0.11,6.19]}$	$5.24_{[2.55,9.03]}$	$16.89_{[7.96,26.50]}$	$53.94_{[50.95,56.81]}$
	Qwen3	$69.10_{[67.24,70.92]}$	$2.56_{[0.14,6.12]}$	$3.19_{[1.37,5.95]}$	$10.98_{[5.13,19.79]}$	$67.45_{[64.64,69.99]}$	$49.89_{[47.55,52.10]}$	$2.34_{[0.14,6.29]}$	$5.15_{[2.75,8.23]}$	$13.00_{[6.26,21.63]}$	$49.11_{[46.48,51.44]}$
	Gemini3	$50.60_{[48.64,52.64]}$	$2.37_{[0.14,6.01]}$	$3.15_{[1.38,5.28]}$	$9.89_{[4.89,17.24]}$	$49.41_{[46.59,51.94]}$	$49.69_{[47.50,51.95]}$	$2.54_{[0.14,6.60]}$	$4.14_{[2.05,6.49]}$	$12.96_{[6.46,20.47]}$	$48.85_{[46.35,51.13]}$

End-to-end bias. We firstly compare the end-to-end fairness of different LLMs across gender and age attributes in Table 1. *All LLMs exhibit varying degrees of unfairness across the two datasets.*[3] In terms of ACC, the Qwen3 model performs exceptionally well on both datasets, achieving 69.10% and 49.89% respectively, while maintaining relatively low ΔACC and DP values, indicating strong fairness. The Qwen3VL also demonstrates high accuracy, but its higher EoD in the Age group suggests some fairness risks in that dimension. Overall, the Qwen series achieves a better balance between fairness and utility (FUT). Note that

[2] Dataset is provided here.

[3] Directly applying Qwen3VL-8B and Gemini3-Flash as end-to-end tools, rather than agents, results in poor overall utility.

DP is closely tied to the base rates of individual subgroups; therefore, careful examination is required before adopting this metric.

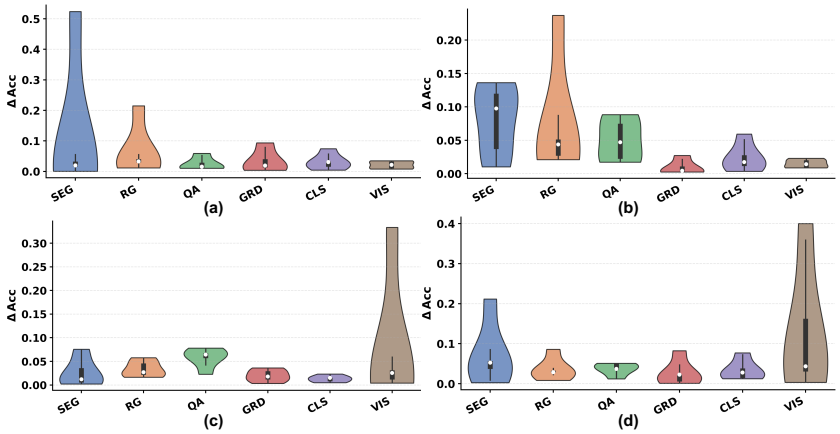

Fig. 2. Tool exposure bias across tools and subgroups. (a) Gender on CheXAgentBench, (b) Age on CheXAgentBench, (c) Gender on MIMIC-FairnessVQA, and (d) Age on MIMIC-FairnessVQA.

Tool-exposure bias. Fig. 2 summarizes *tool exposure* accuracy gaps: for each tool, the violin shows the distribution (across driver LLMs) of $|\Delta\text{ACC}|$ computed on instances where the tool is invoked. Two patterns stand out. *(i) Tool dependence:* on CheXAgentBench, *segmentation* exhibits the largest and heaviest-tailed gaps (especially for gender), with *report generation* a distant second. In contrast, *classifier* and *grounding* remain consistently near zero. *(ii) Dataset and attribute shift:* on MIMIC-FairnessVQA, the largest disparities concentrate in the *visualizer*, suggesting that different tool outputs become the main "fairness bottleneck" under different data distributions. Connecting to the end-to-end results (Table 1), the overall ΔACC gaps are smaller than the worst exposure-conditioned gaps because they average over heterogeneous trajectories and are further shaped by *transition bias* and *LLM reasoning bias*. Nevertheless, the exposure analysis isolates which tools are most likely to seed end-to-end unfairness, providing a mechanistic explanation for why some backbones exhibit larger ΔACC even when overall accuracy is comparable.

Tool-transition bias. As shown in Fig. 3, distinct tool transition patterns are observed across different attributes in various datasets, with gender-based differences being more pronounced than those related to age. In both datasets, female patients are more likely than male patients to proceed directly from the *Classifier* or *Report Generator*. Within the age attribute of CheXAgentBench, younger individuals exhibit a similar tendency compared to older ones. An interesting observation in MIMIC-FairnessVQA is that male patients often call the *Classifier* again after using the *Visualizer*, which is a pattern that warrants fur-

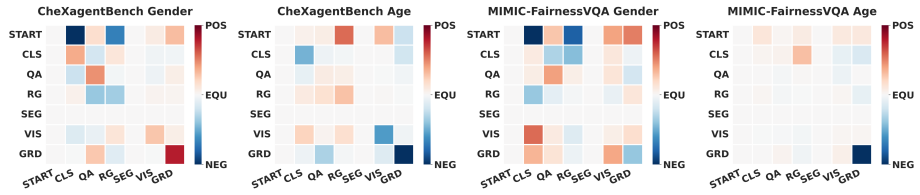

Fig. 3. Tool transition biases represented as matrices. START: start indicator; CLS: classifier; RG: report generator; SEG: segmentator; VIS: visualizer; GRD: phrase grounding. Cells in red denote values larger than zero, while cells in blue denote values smaller than zero.$\Delta_{\mathrm{TTB}}^{\mathrm{gender}} = P^{\mathrm{male}} - P^{\mathrm{female}}$, $\Delta_{\mathrm{TTB}}^{\mathrm{age}} = P^{\mathrm{young}} - P^{\mathrm{old}}$. All subfigures present results averaged across different LLMs.

ther investigation. Additionally, in both the age attribute of CheXAgentBench and MIMIC-FairnessVQA, older individuals and male patients show a higher frequency of repeated calls to *Grounding* tools, which may suggest more effort is needed for answering questions about these demographics.

Table 2. Subgroup gaps of reasoning-bias. $\mathrm{Mean}_{\mathrm{CI}}$ is reported. All values are multiplied by 100 for better readability. Best and Second Best are highlighted.

Configuration		CheXAgentBench			MIMIC-FairnessVQA		
Attr.	LLM	ΔJudgeGap↓	ΔHedge↓	ΔDemo.↓	ΔJudgeGap↓	ΔHedge↓	ΔDemo.↓
Gender	**LLaMA3.1**	$6.19_{[0.21,19.00]}$	$4.97_{[0.25,11.66]}$	$0.09_{[0.02,1.66]}$	$0.10_{[0.30,21.70]}$	$3.10_{[0.20,12.40]}$	$0.10_{[0.00,0.30]}$
	Ministral-3	$2.86_{[0.15,8.92]}$	$12.20_{[0.63,31.73]}$	$0.56_{[0.06,3.74]}$	$3.10_{[0.40,23.40]}$	$4.90_{[0.60,44.40]}$	$0.20_{[0.00,1.20]}$
	Qwen3VL	$4.59_{[0.22,11.39]}$	$534.95_{[3.58,1438.18]}$	$1.57_{[0.08,4.21]}$	$21.67_{[2.66,41.90]}$	$831.10_{[44.75,3489.64]}$	$0.10_{[0.00,0.30]}$
	Qwen3	$5.58_{[0.47,11.69]}$	$12.58_{[0.65,51.20]}$	$3.63_{[0.57,6.88]}$	$1.90_{[0.30,24.80]}$	$4.40_{[1.00,69.20]}$	$0.00_{[0.00,0.00]}$
	Gemini3	$1.67_{[0.09,4.48]}$	$0.89_{[0.13,11.09]}$	$0.01_{[0.06,4.22]}$	$21.20_{[10.49,32.50]}$	$10.40_{[0.80,25.00]}$	$0.30_{[0.00,1.30]}$
Age	**LLaMA3.1**	$2.60_{[0.29,16.73]}$	$5.94_{[0.51,12.88]}$	$1.24_{[0.08,2.71]}$	$22.90_{[4.99,41.11]}$	$1.30_{[0.20,11.10]}$	$0.10_{[0.00,0.30]}$
	Ministral-3	$5.23_{[0.45,11.47]}$	$26.29_{[5.16,50.15]}$	$4.06_{[1.61,6.54]}$	$28.10_{[8.70,47.70]}$	$56.70_{[18.50,94.10]}$	$0.80_{[0.10,1.80]}$
	Qwen3VL	$6.65_{[0.46,13.50]}$	$407.66_{[14.56,1080.75]}$	$3.12_{[1.01,5.53]}$	$15.74_{[1.16,35.15]}$	$1165.50_{[89.77,3893.21]}$	$0.10_{[0.00,0.30]}$
	Qwen3	$1.10_{[0.10,7.75]}$	$4.70_{[0.56,47.52]}$	$2.14_{[0.11,4.94]}$	$24.10_{[4.70,44.60]}$	$67.20_{[9.69,127.11]}$	$0.00_{[0.00,0.00]}$
	Gemini3	$0.14_{[0.05,3.16]}$	$5.88_{[0.27,17.05]}$	$8.26_{[5.08,11.48]}$	$2.40_{[0.30,13.80]}$	$17.40_{[5.00,31.01]}$	$0.50_{[0.00,1.50]}$

LLM-reasoning bias. Table 2 reports subgroup gaps in three response-level features: judge-scored reasoning quality (ΔJudgeGap), hedging frequency (ΔHedge), and explicit demographic mentions (ΔDemo). Two trends are clear. *First,* reasoning bias gaps are highly model-dependent: Qwen3VL shows larger hedging gaps than all other backbones across both datasets and attributes, indicating substantially different uncertainty expression between subgroups. In contrast, LLaMA3.1 and Qwen3 generally exhibit low-to-moderate hedging gaps, while Gemini3 is consistently among the lowest on CheXAgentBench gender. *Second,* the three measures capture distinct behaviors and need not move together: some models achieve low demographic-term gaps (ΔDemo) yet still show large hedging or judge gaps (e.g., Qwen3VL on MIMIC-FairnessVQA), suggesting that subgroup differences primarily manifest in uncertainty style rather than explicit demographic framing. Overall, these results highlight that driver LLMs can intro-

duce substantial subgroup-dependent variation in response synthesis even when the same questions and tool outputs are provided.

4 Conclusion

We audit demographic fairness in MedRAX-style tool-using chest X-ray agents and propose `DUCX` to decompose end-to-end disparities into *tool exposure*, *tool transition*, and *LLM reasoning* biases. Across five driver LLMs and two agent-suitable benchmarks (CheXAgentBench and our MIMIC-FairnessVQA), we find persistent subgroup gaps and show that unfairness can originate from intermediate tool use, routing, and response synthesis, not only final predictions. Our results motivate process-level fairness evaluation for agentic medical imaging systems, and future work will develop stage-targeted mitigation and extend audits to broader clinical tasks and agent designs.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agarwal, S., Yu, H.: Detecting hedge cues and their scope in biomedical text with conditional random fields. Journal of biomedical informatics **43**(6), 953–961 (2010)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
4. Derner, E., De La Fuente, S.S., Gutiérrez, Y., Pozo, P.M., Oliver, N.M.: Leveraging large language models to measure gender representation bias in gendered language corpora. In: Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP). pp. 468–483 (2025)
5. Fallahpour, A., Ma, J., Munim, A., Lyu, H., WANG, B.: MedRAX: Medical reasoning agent for chest x-ray. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=JiFfij5iv0
6. Fallahpour, A., et al.: Chestagentbench. Hugging Face Dataset (2025), https://huggingface.co/datasets/wanglab/chest-agent-bench
7. Glocker, B., Jones, C., Roschewitz, M., Winzeck, S.: Risk of bias in chest radiography deep learning foundation models. Radiology: Artificial Intelligence **5**(6), e230060 (2023)
8. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
9. Hanna, J.J., Wakene, A.D., Johnson, A.O., Lehmann, C.U., Medford, R.J.: Assessing racial and ethnic bias in text generation by large language models for health care–related tasks: Cross-sectional study. Journal of Medical Internet Research **27**, e57257 (2025)

10. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
11. Jackson, N.J., Yan, C., Malin, B.A.: Enhancement of fairness in ai for chest x-ray classification. In: AMIA Annual Symposium Proceedings. vol. 2024, p. 551 (2025)
12. Jin, R., Deng, W., Chen, M., Li, X.: Debiased noise editing on foundation models for fair medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 164–174. Springer (2024)
13. Jin, R., Huang, G., Shen, X., Zhang, Q., Tan, Y.S., Li, X.: See-in-pairs: Reference image-guided comparative vision-language models for medical diagnosis. arXiv preprint arXiv:2506.18140 (2025)
14. Jin, R., Xu, Z., Zhong, Y., Yao, Q., QI, D., Zhou, S.K., Li, X.: Fairmedfm: fairness benchmarking for medical imaging foundation models. Advances in Neural Information Processing Systems **37**, 111318–111357 (2024)
15. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
16. Kim, S.S., Liao, Q.V., Vorvoreanu, M., Ballard, S., Vaughan, J.W.: " i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In: Proceedings of the 2024 ACM conference on fairness, accountability, and transparency. pp. 822–835 (2024)
17. Kim, Y., Park, C., Jeong, H., Chan, Y.S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., Park, H.W.: Mdagents: An adaptive collaboration of llms for medical decision-making. Advances in Neural Information Processing Systems **37**, 79410–79452 (2024)
18. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)
19. Liu, A.H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., et al.: Ministral 3. arXiv preprint arXiv:2601.08584 (2026)
20. Liu, M., Ning, Y., Teixayavong, S., Mertens, M., Xu, J., Ting, D.S.W., Cheng, L.T.E., Ong, J.C.L., Teo, Z.L., Tan, T.F., et al.: A translational perspective towards clinical ai fairness. NPJ digital medicine **6**(1), 172 (2023)
21. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature medicine **27**(12), 2176–2182 (2021)
22. Sharma, N.: Cxr-agent: Vision-language models for chest x-ray interpretation with uncertainty aware radiology reporting. arXiv preprint arXiv:2407.08811 (2024)
23. Wang, P., Lu, W., Lu, C., Zhou, R., Li, M., Qin, L.: Large language model for medical images: A survey of taxonomy, systematic review, and future trends. Big Data Mining and Analytics **8**(2), 496–517 (2025)
24. Wang, W., Ma, Z., Wang, Z., Wu, C., Ji, J., Chen, W., Li, X., Yuan, Y.: A survey of LLM-based agents in medicine: How far are we from baymax? In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025. pp. 10345–10359. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.findings-acl.539, https://aclanthology.org/2025.findings-acl.539/

25. Wu, P., Liu, C., Chen, C., Li, J., Bercea, C.I., Arcucci, R.: Fmbench: Benchmarking fairness in multimodal large language models on medical tasks. arXiv preprint arXiv:2410.01089 (2024)
26. Xu, Z., Li, J., Yao, Q., Li, H., Zhao, M., Zhou, S.K.: Addressing fairness issues in deep learning-based medical image analysis: a systematic review. npj Digital Medicine **7**(1), 286 (2024)
27. Xu, Z., Zhao, S., Quan, Q., Yao, Q., Zhou, S.K.: Fairadabn: Mitigating unfairness with adaptive batch normalization and its application to dermatological disease classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 307–317. Springer (2023)
28. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
29. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)
30. Zhang, H., Dullerud, N., Roth, K., Oakden-Rayner, L., Pfohl, S., Ghassemi, M.: Improving the fairness of chest x-ray classifiers. In: Conference on health, inference, and learning. pp. 204–233. PMLR (2022)
31. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)
32. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., Van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE **109**(5), 820–838 (2021)