MEDSTEER: CLINICALLY CONSTRAINED REPRESENTATION STEERING FOR DEBIASING MEDICAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used for medical decision support and clinical question answering, yet their deployment is hindered by clinically unjustified demographic bias. Mitigating such bias is particularly challenging because demographic attributes like age, gender, and ethnicity may be relevant in certain contexts while introducing inappropriate sensitivity in others. Existing steering methods often fail to separate sensitive attributes from diagnosis-relevant clinical information. To bridge this gap, we propose MedSteer, a clinically constrained, inference-time representation steering framework tailored for medical LLMs. MedSteer isolates demographic bias by estimating representation shifts from diagnosis-conditioned matched pairs, ensuring that steering strictly targets ungrounded sensitivity without disrupting valid diagnostic pathways. To determine optimal steering interventions, MedSteer introduces a novel prioritization strategy based on semantic stability and clinical orthogonality, effectively neutralizing non-decisive demographic bias while safeguarding core clinical reasoning. Furthermore, it integrates a parameter-free auxiliary lexical control signal derived from stigma–neutral contrasts to further mitigate stylistic and lexical bias. Evaluated across four medical LLMs on sensitivity analysis, long-text entity retention, and medical question answering, MedSteer outperforms an unweighted Mean-Diff baseline. It improves average question answering accuracy by 2.6 percentage points across 12 model–dataset combinations, better retains clinical entities, and achieves a superior fairness–utility trade-off. These findings highlight MedSteer as a practical post-deployment solution to mitigate demographic bias in medical LLMs without compromising clinical performance.

1 INTRODUCTION

Large language models (LLMs) are increasingly being considered for clinical question answering, patient communication, decision support, and medical education. However, their responses may change when a patient's demographic or social descriptors are altered, even when the clinically relevant evidence remains unchanged. When unsupported by the available evidence, such sensitivity can affect diagnostic hypotheses, risk communication, and treatment-related recommendations, potentially reproducing or amplifying existing disparities in healthcare.

Clinical debiasing, however, cannot be reduced to making a model invariant to all sensitive attributes. Age can affect disease prevalence and treatment eligibility; sex and gender can influence screening recommendations and adverse-event profiles; and ethnicity-associated social context may reflect differences in exposure, access to care, or structural inequities (Pendse et al., 2026; Templin et al., 2025; Hanlon et al., 2025; Ji et al., 2025a; Zhang et al., 2025). Conversely, these attributes can also be misused as biological proxies or can induce different recommendations when the medically relevant evidence is otherwise unchanged (Omar et al., 2025; Ji et al., 2025a; Templin et al., 2025). We refer to an attribute as *clinically non-decisive* for a case when changing it, while holding the relevant clinical evidence fixed, should not alter the medically justified conclusion. The desired behavior is therefore not unconditional demographic invariance, but *clinically conditional invariance*: a model should suppress the influence of a sensitive attribute when it is non-decisive, while preserving that information when it is required for valid clinical reasoning. This distinction is particularly important

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for ethnicity, for which clinically relevant social context and inappropriate biological essentialism can be difficult to disentangle (Fiske et al., 2025).

The central technical challenge is representation entanglement. Within an LLM's hidden states, demographic cues may be encoded jointly with diagnosis, symptom severity, social context, and lexical style. Consequently, an attribute direction estimated from generic contrasts may contain both undesirable demographic associations and clinically meaningful information. Steering along such a direction can fail in two opposite ways: an insufficient intervention may leave harmful sensitivity largely unchanged, whereas an aggressive intervention may suppress evidence needed for accurate medical reasoning.

Existing research addresses complementary parts of this problem. General-domain natural language processing has studied both training-time behavior alignment (Li et al., 2024; Hall et al., 2025) and inference-time representation steering, which modifies hidden states during generation without updating model parameters (Siddique et al., 2026; Luo et al., 2024; Qiu et al., 2024). In parallel, medical LLM research has introduced benchmarks for measuring demographic sensitivity, counterfactual consistency, and fairness-related performance across clinical tasks (Lamparth et al., 2026; Pfohl et al., 2024; Rawat et al., 2024; Ji et al., 2025b; Ghosh et al., 2025). These two lines of work, however, remain only loosely connected. Audit-oriented benchmarks characterize when demographic sensitivity occurs but do not directly intervene during generation. Conversely, representation-steering methods can alter model behavior at inference time, but many construct attribute directions from unconditioned or uniformly aggregated contrasts without explicitly examining their overlap with diagnosis-relevant representations. The missing capability is therefore a *selective intervention*: identifying which components of an attribute-associated representation shift can be attenuated without perturbing the clinical evidence underlying the model's response.

To address this challenge, we introduce MEDSTEER, a diagnosis-aware inference-time representation intervention framework for clinically constrained debiasing. MEDSTEER first estimates attribute-associated representation shifts from diagnosis-conditioned matched pairs, such that candidate directions are derived under comparable clinical contexts rather than across potentially confounded examples. It then weights these candidate directions according to two complementary criteria: *semantic stability*, which favors attribute-associated shifts that remain consistent across matched contexts, and *clinical orthogonality*, which downweights shifts that are strongly coupled with diagnosis-specific information. Finally, an auxiliary control direction derived from stigma–neutral lexical contrasts complements demographic steering by capturing undesirable associations that may not be fully expressed through demographic contrasts alone. The resulting intervention operates entirely at inference time and leaves the model parameters unchanged. Rather than attempting to erase the complete representation of a sensitive attribute, MEDSTEER targets residual attribute-associated variation after conditioning on the clinical context.

We evaluate MEDSTEER across four medical LLMs using demographic-sensitivity assessments, long-form clinical entity retention, and medical question answering. We use unweighted MeanDiff as the primary controlled steering baseline because it follows the same difference-based intervention paradigm while replacing diagnosis-aware weighting with uniform direction averaging. Relative to MeanDiff, MEDSTEER achieves higher entity retention in long-form clinical generations, improves average medical question-answering accuracy by 2.6 percentage points across 12 model–dataset combinations, and yields a more favorable fairness–utility trade-off. These results indicate that conditioning representation steering on clinical context can reduce non-decisive demographic sensitivity without sacrificing, and in some settings while improving, downstream medical utility.

The main contributions are as follows:

- We formulate medical LLM debiasing as a clinically constrained inference-time control problem under a conditional-invariance principle: demographic effects should be suppressed when they are clinically non-decisive, but preserved when they support valid medical reasoning.
- We propose MEDSTEER, a diagnosis-aware representation-steering framework that estimates attribute-associated shifts from matched clinical contexts and selectively aggregates them using semantic stability, clinical orthogonality, and an auxiliary stigma–neutral lexical control signal.

- We conduct systematic evaluations across four medical LLMs and multiple downstream tasks, showing that MedSteer improves the fairness–utility trade-off over unweighted MeanDiff steering, with better long-text entity retention and maintained or improved medical question answering performance.

Taken together, this study frames medical debiasing not as the removal of all demographic information, but as selective control under constraints that protect clinical meaning. This framing offers a practical path toward safer and more trustworthy deployment of medical LLMs.

2 Related Work

2.1 Bias, Fairness, and Evaluation in Medical LLMs

A growing body of literature shows that medical LLMs can be sensitive to sociodemographic cues across clinical tasks, from diagnosis to decision support (Benkirane et al., 2025; Ahsan et al., 2025; Liu et al., 2025; Omar et al., 2024; Levartovsky et al., 2025; Bouguettaya et al., 2025). Benchmarking efforts such as EquityMedQA and DiversityMedQA have established counterfactual evaluation protocols, which test whether model outputs change when patient attributes are varied while the clinical case is otherwise held fixed (Pfohl et al., 2024; Rawat et al., 2024; Ji et al., 2025b; Ghosh et al., 2025). More recent evaluations extend these protocols to clinician-annotated mental-health decisions and to controlled perturbations of demographics and vital signs in ICU notes (Lamparth et al., 2026; Tagliabue et al., 2026). A critical gap remains: current methods largely focus on identifying disparities, while fewer approaches provide actionable, clinically sensitive interventions for real-time generation (Poulain et al., 2024; Suenghataiphorn et al., 2025).

2.2 Stigmatizing Language in Clinical NLP and Medical LLMs

Research on representational harm has identified stigmatizing language as a recurring problem in both electronic health records (EHRs) and model-generated clinical text (Himmelstein et al., 2022; Hulchafo et al., 2025; Bradford et al., 2024; Volkow et al., 2021; Werder et al., 2022; Harrigian et al., 2023). Recent studies further show that LLMs may reproduce or amplify such language in high-stakes areas, including addiction care and mental health (Sethi et al., 2026; Chen et al., 2025; Scroggins et al., 2025; Wang et al., 2026; Moore et al., 2025). However, this literature often treats lexical stigmatization as a separate wording problem, rather than as part of a broader pattern of demographic and social sensitivity in medical LLMs.

2.3 Inference-Time Steering and Representation Engineering

Methodologically, MedSteer is related to representation engineering and activation steering, which adjust internal model representations at test time (Qiu et al., 2024; Luo et al., 2024; Kong et al., 2024). Prior work has shown that training-free mechanisms can modify model behavior without changing model parameters (Konen et al., 2024; Siddique et al., 2026; Im & Li, 2025; Wehner et al., 2025). Other clinical activation-steering studies target sycophantic reasoning or hallucination, and show the value of context-dependent or gated intervention while exposing trade-offs with clinical knowledge retention (Ayadi & Hautli, 2026; Tripathi et al., 2026). These studies motivate selective intervention, but do not establish that a demographic steering direction is safe to apply across diagnoses: it may contain both unwanted associations and clinically relevant information. MedSteer addresses this limitation by weighting candidate directions according to semantic stability and clinical orthogonality before intervention.

3 Method

3.1 Problem Formulation

Let f denote a frozen medical language model, and let $h_{\ell,t}(x) \in \mathbb{R}^m$ denote its hidden representation at layer ℓ and token position t for input x. For direction estimation, we use the representation at the final prompt token, denoted by $h_\ell(x) \equiv h_{\ell,T}(x)$.

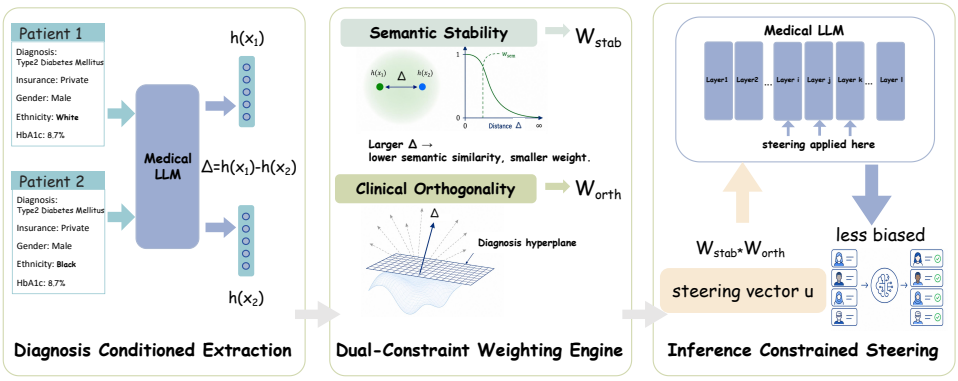


Figure 1: Overview of the MedSteer framework. Diagnosis-conditioned extraction constructs matched clinical pairs that share the same diagnosis but differ in a target demographic or social attribute, and extracts the corresponding hidden-state difference. The dual-constraint weighting engine scores each candidate direction by semantic stability and clinical orthogonality. During inference, the optimized steering vector is injected into selected layers of a frozen medical LLM.

For a demographic or social attribute $a \in \mathcal{A}$, let $\mathcal{T}_{a:u\to v}(x)$ denote an attribute intervention that changes a from value u to v while holding the task-relevant clinical evidence fixed. Following the conditional-invariance principle, we call a *clinically non-decisive* for x when this intervention should not change the medically justified conclusion. Our objective is therefore not to erase all information associated with a, but to reduce attribute-associated sensitivity that is weakly coupled with the clinical evidence required for valid reasoning.

Conceptually, we seek an intervention direction u that reduces attribute sensitivity while preserving medical utility:

$$\min_{u} \ \mathcal{B}(u) \qquad \text{s.t.} \qquad \mathcal{U}(u) \geq \mathcal{U}(0) - \epsilon, \tag{1}$$

where $\mathcal{B}$ denotes sensitivity to clinically non-decisive attribute variation and $\mathcal{U}$ denotes downstream medical utility.

We use Eq. (1) as a design principle rather than optimizing $\mathcal{B}$ and $\mathcal{U}$ directly. MedSteer approximates the utility-preservation constraint through diagnosis-conditioned contrasts and two representation-level constraints: semantic stability and clinical orthogonality.

As illustrated in Figure 1, the resulting pipeline consists of diagnosis-conditioned direction extraction, dual-constraint weighting, and inference-time representation steering. Stigma–neutral lexical contrasts are treated separately as an auxiliary control source and are introduced after the demographic direction is estimated.

3.2 Diagnosis-Conditioned Candidate Direction Extraction

We construct diagnosis-conditioned matched pairs from MIMIC-III (Johnson et al., 2016). For diagnosis $c \in \mathcal{C}$ and target attribute $a \in \mathcal{A}$, a pair $(x_{i,1}^{(a,c)}, x_{i,2}^{(a,c)})$ shares the same diagnosis and is matched on the remaining observed demographic and social covariates, while differing in the target attribute a. These pairs therefore provide approximate counterfactual-style contrasts under comparable clinical contexts. Detailed preprocessing and matching criteria are provided in Appendix C.1 and C.2.

At layer ℓ, we compute the sample-level attribute-associated shift

$$\Delta h_{i,c}^{(a,\ell)} = h_\ell\left(x_{i,1}^{(a,c)}\right) - h_\ell\left(x_{i,2}^{(a,c)}\right). \tag{2}$$

Each $\Delta h_{i,c}^{(a,\ell)}$ is a candidate direction that may contain both attribute-associated variation and residual clinical variation. The weighting stage below reduces the contribution of candidates that are poorly matched semantically or strongly aligned with diagnosis-level clinical content.

Not all matched-pair differences provide equally selective attribute signals. MedSteer assigns each candidate direction a diagnosis-aware weight using two complementary criteria.

3.2.1 SEMANTIC STABILITY

We first quantify whether the matched pair remains close in representation space after changing the target attribute:

$$w_{\mathrm{stab},i}^{(\ell)} = \max\Big(0,\, \cos\Big(h_\ell(x_{i,1}^{(a,c)}), h_\ell(x_{i,2}^{(a,c)})\Big)\Big)\,. \tag{3}$$

A larger value indicates that the shared clinical context remains representation-consistent despite the attribute perturbation, making $\Delta h_{i,c}^{(a,\ell)}$ less likely to reflect broad semantic drift.

3.2.2 CLINICAL ORTHOGONALITY

We next penalize candidate directions that align strongly with diagnosis-level representations. Let $\mathcal{P}_c$ denote all matched pairs associated with diagnosis c. We define the diagnosis anchor as the centroid of their representations:

$$v_c^{(\ell)} = \frac{1}{2|\mathcal{P}_c|} \sum_{(i,a)\in\mathcal{P}_c} \Big[h_\ell(x_{i,1}^{(a,c)}) + h_\ell(x_{i,2}^{(a,c)})\Big]\,. \tag{4}$$

The corresponding alignment penalty is

$$w_{\mathrm{orth},i}^{(\ell)} = 1 - \Big|\cos\Big(\Delta h_{i,c}^{(a,\ell)}, v_c^{(\ell)}\Big)\Big|\,. \tag{5}$$

Thus, candidate shifts more nearly orthogonal to the diagnosis anchor receive larger weights.

We combine the two constraints multiplicatively:

$$w_i^{(\ell)} = w_{\mathrm{stab},i}^{(\ell)}\, w_{\mathrm{orth},i}^{(\ell)}. \tag{6}$$

Let $\mathcal{P}_a$ denote all diagnosis-conditioned pairs constructed for attribute a. The attribute-specific demographic direction is

$$u_a^{(\ell)} = \frac{\sum_{(i,c)\in\mathcal{P}_a} w_i^{(\ell)} \Delta h_{i,c}^{(a,\ell)}}{\sum_{(i,c)\in\mathcal{P}_a} w_i^{(\ell)}}. \tag{7}$$

Compared with uniform MeanDiff aggregation, Eq. (7) reduces the influence of candidate shifts that either disturb the shared clinical representation or overlap strongly with diagnosis-level content.

3.3 AUXILIARY LEXICAL CONTROL

MedSteer additionally uses stigma–neutral lexical contrasts as an auxiliary framing-sensitive signal. Let $\mathcal{P}_{\mathrm{lex}} = \{(s_j^{\mathrm{stig}}, s_j^{\mathrm{neu}})\}_{j=1}^{N_{\mathrm{lex}}}$ denote the lexical pairs, where both expressions are inserted into the same clinical template $T_j(\cdot)$. Their layer-wise difference is

$$\Delta h_{j,\mathrm{lex}}^{(\ell)} = h_\ell\Big(T_j(s_j^{\mathrm{stig}})\Big) - h_\ell\big(T_j(s_j^{\mathrm{neu}})\big)\,. \tag{8}$$

Because this branch provides lexical rather than diagnosis-specific supervision, we apply only the semantic-stability criterion:

$$\begin{aligned} w_{j,\mathrm{lex}}^{(\ell)} &= \max\Big(0,\, \cos\Big(h_\ell(T_j(s_j^{\mathrm{stig}})), h_\ell(T_j(s_j^{\mathrm{neu}}))\Big)\Big)\,, \\ u_{\mathrm{lex}}^{(\ell)} &= \frac{\sum_j w_{j,\mathrm{lex}}^{(\ell)} \Delta h_{j,\mathrm{lex}}^{(\ell)}}{\sum_j w_{j,\mathrm{lex}}^{(\ell)}}. \end{aligned} \tag{9}$$

Let $u_{\mathrm{demo}}^{(\ell)}$ denote the demographic direction used for a given intervention. For attribute-specific steering, $u_{\mathrm{demo}}^{(\ell)} = u_a^{(\ell)}$. The demographic and lexical signals are then combined as

$$\widetilde{u}^{(\ell)} = \mathrm{Norm}\Big(u_{\mathrm{demo}}^{(\ell)} + u_{\mathrm{lex}}^{(\ell)}\Big)\,. \tag{10}$$

Here, $\mathrm{Norm}(\cdot)$ denotes the normalization used in the implementation to place the combined direction on a consistent scale.

3.4 Inference-Time Representation Steering

After estimating $\widetilde{u}^{(\ell)}$, MedSteer intervenes directly in the hidden states of the frozen model. During autoregressive generation, for each selected layer $\ell \in \mathcal{L}$, the hidden representation is updated as

$$\widehat{h}_{\ell,t} = h_{\ell,t} - \alpha\, \widetilde{u}^{(\ell)}, \qquad \ell \in \mathcal{L}, \tag{11}$$

where α controls intervention strength. No model parameters are updated. The diagnosis conditioning and dual-constraint weighting are used when constructing the steering direction; MedSteer does not require a separate test-time classifier for determining clinical decisiveness.

The magnitude of α determines the fairness–utility trade-off: weak steering may leave residual attribute sensitivity, whereas overly strong steering may perturb medically useful representations. We therefore select α using the same tuning protocol for all steering-based methods and report the model-specific settings in Appendix B.

4 Experiments

4.1 Datasets and Benchmarks

For direction estimation, we use a structured subset of MIMIC-III (Johnson et al., 2016) together with an auxiliary stigma–neutral lexical resource. MIMIC-III is a publicly available, de-identified critical care database, and all use of the data in this study follows the database access requirements. Access requires completion of the prescribed data-use training and agreement to the corresponding data-use terms. From MIMIC-III (Johnson et al., 2016), we construct diagnosis-conditioned matched pairs by varying a single target attribute while matching the remaining normalized demographic and social covariates. This yields 13,264 matched pairs in total, covering ethnicity, gender, insurance, marital status, and religion. Figure 4 summarizes the scale and coverage of the data used for direction estimation. The matched-pair pool is distributed across five demographic or social dimensions, while the auxiliary stigma–neutral lexical resource provides balanced lexical contrasts for the supplementary control branch. The prompt template and extraction details are provided in Appendix A and Appendix C.3.

Medical utility is measured on MedQA, MedMCQA, and four medically relevant MMLU subsets. To assess long-text entity retention, we prompt models to rewrite MIMIC clinical records and evaluate entity retention with an NER-based F1 score. Fairness is evaluated with EquityMedQA for counterfactual sensitivity analysis and DiversityMedQA for subgroup-level disparity analysis.

4.2 Models and Baselines

We evaluate four open-source backbones: Meditron3-8B (Sallinen et al., 2025), Internist-7B (Griot et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), and Meerkat-7B (Kim et al., 2025). We compare MedSteer against two baselines: Base, the unmodified model, and MeanDiff, an unweighted mean-difference steering baseline that averages hidden-state differences without semantic-stability or clinical-orthogonality weighting. MeanDiff is the principal steering baseline because it directly tests whether the proposed diagnosis-aware weighting improves over unweighted direction averaging. All steering-based methods use the same prompt format, decoding configuration, and hyperparameter search protocol.

4.3 Results and Analysis

This subsection reports the empirical evidence for the experimental goals defined above: preserving medical utility and reducing fairness-related sensitivity. We interpret the results conservatively, especially when the trends differ across model backbones.

4.3.1 Medical Utility and Clinical Fidelity

Long-text entity retention: Table 1 reports entity-level F1 on MIMIC clinical note rewriting. Across all four backbones, MedSteer achieves the highest reported F1 among the compared methods. This result is important because long-text entity retention is vulnerable to overly aggressive steering:

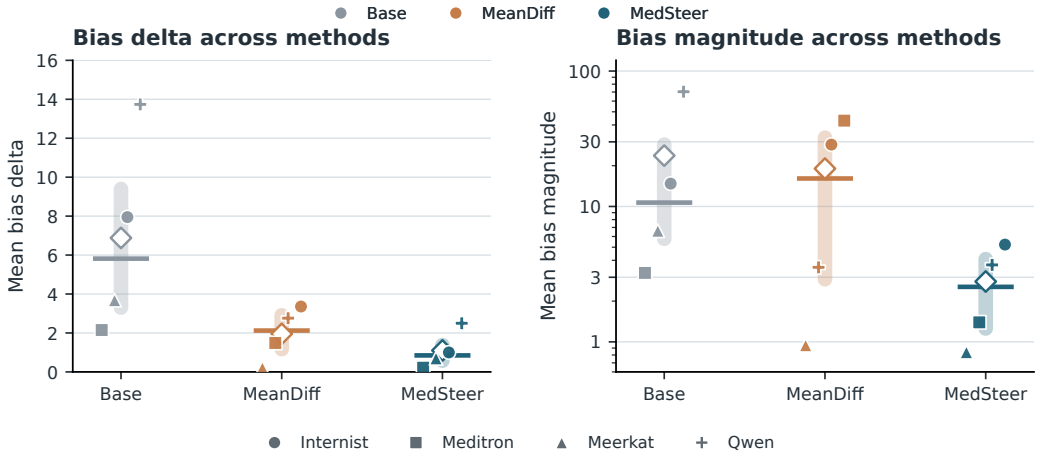


Figure 2: Projection-based bias metrics on EquityMedQA. Lower values indicate smaller residual sensitivity along the evaluation bias direction. Left: Mean Bias Delta measures the average projection difference between matched evaluation pairs along the bias direction. Right: Mean Bias Magnitude measures the average absolute projection magnitude on the same direction after intervention.

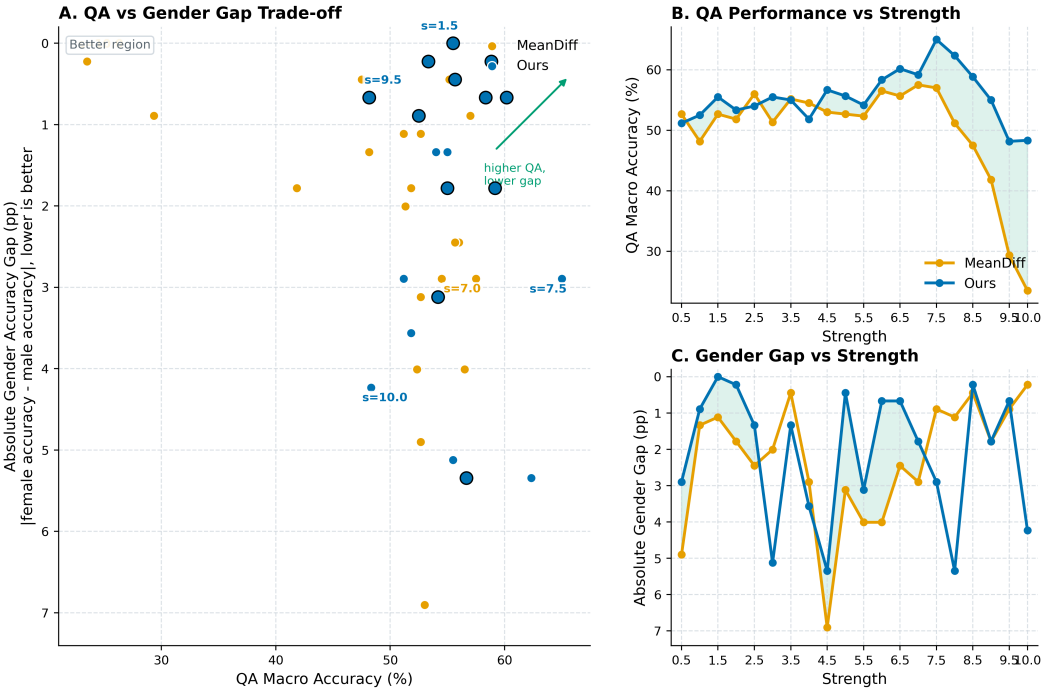


Figure 3: Fairness–utility behavior across steering strengths in Meditron3-8B. Panel (A) plots QA macro accuracy against the absolute gender gap. Panels (B) and (C) show QA macro accuracy and the absolute gender gap as the steering strength varies. Lower gender gap is better.

Table 1: Entity retention and medical question answering accuracy.

Model	Method	F1 ↑	MedQA ↑	MedMCQA ↑	MMLU ↑
Internist-7B	Base	0.6157	0.5000	0.4564	0.5498
	MeanDiff	0.5766	0.4950	0.4470	0.5241
	MedSteer	**0.6552**	0.5050	**0.4823**	**0.5707**
	MedSteer w/o stigma	0.6285	**0.5195**	0.4677	0.5603
	Δ (MS–MD)	*+0.0786*	*+0.0100*	*+0.0353*	*+0.0466*
Qwen2.5-7B	Base	0.5302	0.6394	0.6080	0.7629
	MeanDiff	0.5299	0.6084	0.5444	0.7191
	MedSteer	**0.5752**	**0.6384**	0.5844	0.7380
	MedSteer w/o stigma	0.4049	0.6280	**0.5969**	**0.7422**
	Δ (MS–MD)	*+0.0453*	*+0.0300*	*+0.0400*	*+0.0189*
Meerkat-7B	Base	0.5302	0.7054	0.5742	0.5842
	MeanDiff	0.6160	0.7106	**0.5813**	0.5756
	MedSteer	**0.6194**	**0.7200**	0.5600	0.6200
	MedSteer w/o stigma	0.6005	0.7045	0.5685	**0.6207**
	Δ (MS–MD)	*+0.0034*	*+0.0094*	*-0.0213*	*+0.0444*
Meditron3-8B	Base	0.5804	0.6974	0.4038	0.5954
	MeanDiff	0.6707	0.6890	0.4588	0.6200
	MedSteer	**0.6811**	0.7165	0.4925	0.6538
	MedSteer w/o stigma	0.6612	**0.7320**	**0.5197**	**0.6621**
	Δ (MS–MD)	*+0.0104*	*+0.0275*	*+0.0337*	*+0.0338*

Note: F1 is the entity-level retention score for rewritten MIMIC clinical notes. MedQA, MedMCQA, and MMLU report medical question answering accuracy. MS and MD denote MedSteer and MeanDiff, respectively. Delta rows report MedSteer minus MeanDiff; higher values are better.

Table 2: Subgroup performance and disparity across gender and ethnicity.

Model	Method	Gender			Ethnicity	
		Male	Female	Gap ↓	Max–Min ↓	CV ↓
Internist-7B	MeanDiff	0.4632	0.4410	0.0222	0.0651	0.0529
	MedSteer	0.4655	0.4499	**0.0156**	**0.0323**	0.0335
	MedSteer w/o stigma	0.4855	0.5078	0.0223	0.0370	**0.0294**
	Δ (MS–MD)	*+0.0023*	*+0.0089*	*-0.0066*	*-0.0328*	*-0.0194*
Qwen2.5-7B	MeanDiff	0.7817	0.7728	**0.0089**	0.0264	0.0138
	MedSteer	0.7506	0.7639	0.0134	**0.0141**	**0.0080**
	MedSteer w/o stigma	0.5791	0.6080	0.0290	0.0158	0.0101
	Δ (MS–MD)	*-0.0311*	*-0.0089*	*+0.0045*	*-0.0123*	*-0.0058*
Meerkat-7B	MeanDiff	0.6147	0.5880	0.0267	0.0458	0.0304
	MedSteer	0.5702	0.5835	**0.0134**	**0.0264**	**0.0172**
	MedSteer w/o stigma	0.5457	0.6236	0.0780	0.0405	0.0232
	Δ (MS–MD)	*-0.0445*	*-0.0045*	*-0.0133*	*-0.0194*	*-0.0132*
Meditron3-8B	MeanDiff	0.4543	0.4254	0.0290	0.0423	0.0311
	MedSteer	0.4855	0.4878	**0.0022**	**0.0335**	**0.0246**
	MedSteer w/o stigma	0.5301	0.5212	0.0089	0.0370	0.0294
	Δ (MS–MD)	*+0.0312*	*+0.0624*	*-0.0268*	*-0.0088*	*-0.0065*

Note: For gender, Gap denotes the absolute male–female accuracy gap. For ethnicity, Max–Min is the maximum minus minimum accuracy across five ethnicity variants, and CV is the coefficient of variation. MS and MD denote MedSteer and MeanDiff, respectively. Lower Gap, Max–Min, and CV indicate smaller subgroup disparity.

if bias-aligned and clinically relevant features are insufficiently disentangled, an intervention may suppress medically important content together with unwanted attribute-sensitive variation. The observed gains support the view that diagnosis-aware weighted steering is more selective than unweighted mean-difference steering in this setting.

Medical QA utility: Table 1 shows results on MedQA, MedMCQA, and MMLU. Averaged across all 12 model-dataset combinations, MedSteer improves over MeanDiff by approximately 2.6 percentage points in observed medical QA accuracy. Relative to the Base model, however, the pattern is mixed: MedSteer improves upon Base in some model-benchmark pairs while remaining slightly below Base in others. We therefore interpret these results conservatively. The proposed weighting scheme improves the fairness–utility trade-off relative to unweighted MeanDiff steering while largely preserving downstream medical reasoning.

Role of the auxiliary lexical signal: Ablations show that the lexical branch is not a generic QA booster: its effect on multiple-choice accuracy is mixed, while it consistently improves long-text clinical rewriting and generally yields a better fairness profile. We therefore view it as an auxiliary bias-control signal that improves the overall fairness–fidelity trade-off, which QA accuracy alone does not fully capture. Figure 5 provides a qualitative example on Qwen2.5-7B, illustrating how the lexical steering coefficient changes response framing in ways that multiple-choice QA accuracy does not capture. Please refer to the detailed extraction information in Appendices C.4 and D .

Together, Table 1 and Table 2 indicate that MedSteer is more selective than unweighted MeanDiff steering in the evaluated settings.

4.3.2 Fairness and Hyperparameter Robustness

Representation-level invariance on EquityMedQA: Figure 2 reports two projection-based metrics on EquityMedQA: Mean Bias Delta and Mean Bias Magnitude, where lower values indicate smaller residual sensitivity along the evaluation bias direction. Compared with Base and MeanDiff, MedSteer yields lower values and a more compact distribution on both metrics. These results provide representation-level evidence that weighted steering is associated with lower residual bias-aligned sensitivity than the unweighted baseline.

Hyperparameter Robustness: Figure 3 shows how the fairness–utility trade-off changes with steering strength for Meditron3-8B. MedSteer generally achieves higher QA accuracy than MeanDiff at the same strength, particularly under stronger interventions. However, the gender gap fluctuates for both methods rather than decreasing consistently as steering strength increases. These results motivate selecting the steering coefficient using both utility and fairness on the validation set, rather than assuming that stronger steering improves both.

5 Conclusion

We introduced MedSteer, a clinically constrained inference-time representation-steering framework for reducing non-decisive demographic sensitivity in medical LLMs. By estimating attribute-associated shifts from diagnosis-conditioned matched pairs and weighting them according to semantic stability and clinical orthogonality, MedSteer targets residual attribute sensitivity while limiting interference with diagnosis-relevant representations. An auxiliary stigma–neutral lexical direction further extends the control signal without parameter updates. Across four medical LLMs, MedSteer yields a more favorable fairness–utility trade-off than unweighted MeanDiff steering, with better clinical entity retention and an average 2.6 percentage-point improvement in medical QA accuracy over MeanDiff across 12 model–dataset combinations. These results support diagnosis-aware representation steering as a practical post-deployment approach to clinically constrained debiasing.

AI use statement

We used generative AI tools to assist with language polishing, improving the clarity of parts of the manuscript, and LaTeX formatting and presentation. Generative AI was not used for research design, experiments, data analysis, or scientific conclusions. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content.

Ethics statement

This study uses de-identified MIMIC-III data under its credentialed-access and data-use requirements, with no new human-subject data collection. MedSteer is intended for research on medical-

LLM bias mitigation rather than direct clinical use, and further validation is required before clinical deployment.

Reproducibility statement

The datasets analyzed during the current study, including MIMIC-III, are available in the PhysioNet repository (`https://physionet.org/about/database/`). Note that access to these datasets requires credentialed access and a signed data use agreement. The custom code and scripts used for data processing and analysis in this study are openly available in an anonymized repository at `https://anonymous.4open.science/r/MedSteer-8A3C/`.

References

Hiba Ahsan, Arnab Sen Sharma, Silvio Amir, David Bau, and Byron C Wallace. Elucidating mechanisms of demographic bias in llms for healthcare. *arXiv preprint arXiv:2502.13319*, 2025.

Bayram Ayadi and Annette Hautli. Probing bias formation in medical llms through activation steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 609–620, 2026.

Kenza Benkirane, Jackie Kay, and Maria Perez-Ortiz. How can we diagnose and treat bias in large language models for clinical decision-making? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2263–2288, 2025.

Ayoub Bouguettaya, Elizabeth M Stuart, and Elias Aboujaoude. Racial bias in ai-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. *NPJ Digital Medicine*, 8(1):332, 2025.

William S Bradford, Reed WR Bratches, Hollie Porras, David R Chen, Kelly W Gagnon, and Simon B Ascher. Occurrence of stigmatizing documentation among hospital medicine encounters with opioid-related diagnosis codes: cohort study. *JMIR Formative Research*, 8(1):e53510, 2024.

Hongbo Chen, Myrtede Alfred, and Eldan Cohen. Efficient detection of stigmatizing language in electronic health records via in-context learning: comparative analysis and validation study. *JMIR Medical Informatics*, 13(1):e68955, 2025.

Amelia Fiske, Sarah Blacker, Lester Darryl Genevieve, Theresa Willem, Marie-Christine Fritzsche, Alena Buyx, Leo Anthony Celi, and Stuart McLennan. Weighing the benefits and risks of collecting race and ethnicity data in clinical settings for medical artificial intelligence. *The Lancet Digital Health*, 7(4):e286–e294, 2025.

Rajarshi Ghosh, Abhay Gupta, Hudson McBride, Anurag Jayant Vaidya, and Faisal Mahmood. Medequalqa: Evaluating biases in llms with counterfactual reasoning. In *Proceedings of The First Workshop on Human–LLM Collaboration for Ethical and Responsible Science Production (SciProdLLM)*, pp. 25–37, 2025.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Impact of high-quality, mixed-domain data on the performance of medical language models. *Journal of the American Medical Informatics Association*, 31(9):1875–1883, 2024.

Zara Hall, Melanie Subbiah, Thomas P Zollo, Kathleen McKeown, and Richard Zemel. Guiding llm decision-making with fairness reward models. *arXiv preprint arXiv:2507.11344*, 2025.

Peter Hanlon, Elaine Butterly, Lili Wei, Heather Wightman, Saleh Ali M Almazam, Khalid Alsallumi, Jamie Crowther, Ryan McChrystal, Heidi Rennison, Katherine Hughes, et al. Age and sex differences in efficacy of treatments for type 2 diabetes: a network meta-analysis. *Jama*, 333(12): 1062–1073, 2025.

Keith Harrigian, Ayah Zirikly, Brant Chee, Alya Ahmad, Anne Links, Somnath Saha, Mary Catherine Beach, and Mark Dredze. Characterization of stigmatizing language in medical records. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 2: Short papers)*, pp. 312–329, 2023.

Gracie Himmelstein, David Bates, and Li Zhou. Examination of stigmatizing language in the electronic health record. *JAMA network open*, 5(1):e2144967, 2022.

Ismael Ibrahim Hulchafo, Jihye Kim Scroggins, Sarah E Harkins, Hans Moen, Michele Tadiello, Kenrick Cato, Anahita Davoudi, Dena Goffman, Janice James Aubey, Coretta Green, et al. Stigmatizing and positive language in birth clinical notes associated with race and ethnicity. *JAMA Network Open*, 8(5):e259599, 2025.

Shawn Im and Sharon Li. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*, 2025.

Yuelyu Ji, Wenhe Ma, Sonish Sivarajkumar, Hang Zhang, Eugene M Sadhu, Zhuochun Li, Xizhi Wu, Shyam Visweswaran, and Yanshan Wang. Mitigating the risk of health inequity exacerbated by large language models. *npj Digital Medicine*, 8(1):246, 2025a.

Yuelyu Ji, Hang Zhang, and Yanshan Wang. Bias evaluation and mitigation in retrieval-augmented medical question-answering systems. *arXiv preprint arXiv:2503.15454*, 2025b.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, et al. Small language models learn enhanced reasoning skills from medical textbooks. *NPJ digital medicine*, 8(1):240, 2025.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 782–802, 2024.

Lingkai Kong, Haorui Wang, Wenhao Mu, Yuanqi Du, Yuchen Zhuang, Yifei Zhou, Yue Song, Rongzhi Zhang, Kai Wang, and Chao Zhang. Aligning large language models with representation editing: A control perspective. *Advances in Neural Information Processing Systems*, 37:37356–37384, 2024.

Max Lamparth, Declan Grabb, Amy Franks, Scott Gershan, Kaitlyn Kunstman, Aaron Lulla, Monika Roots, Manu Sharma, Aryan Shrivastava, Nina Vasan, et al. Moving beyond medical exams: A clinician-annotated fairness dataset of real-world tasks and ambiguity in mental healthcare. In *International Conference on Learning Representations*, volume 2026, pp. 23894–23939, 2026.

Asaf Levartovsky, Mahmud Omar, Girish N Nadkarni, Uri Kopylov, and Eyal Klang. Sociodemographic bias in large language model–assisted gastroenterology. *JAMA Network Open*, 8(9): e2532692, 2025.

Jiahui Li, Hanlin Zhang, Fengda Zhang, Tai-Wei Chang, Kun Kuang, Long Chen, and Jun Zhou. Optimizing language models with fair and stable reward composition in reinforcement learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10122–10140, 2024.

Chenxi Liu, Jianing Zheng, Yushu Liu, Xi Wang, Yuting Zhang, Qiang Fu, Wenwen Yu, Ting Yu, Wang Jiang, Dan Wang, et al. Potential to perpetuate social biases in health care by chinese large language models: a model evaluation study. *International Journal for Equity in Health*, 24(1): 206, 2025.

Jinqi Luo, Tianjiao Ding, Kwan H Chan, Darshan Thaker, Aditya Chattopadhyay, Chris Callison-Burch, and René Vidal. Pace: Parsimonious concept engineering for large language models. *Advances in Neural Information Processing Systems*, 37:99347–99381, 2024.

Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C Ong, and Nick Haber. Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 599–627, 2025.

Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U Apakama, Carol R Horowitz, Alexander W Charney, Robert Freeman, Benjamin Kummer, Benjamin S Glicksberg, et al. Socio-demographic biases in medical decision-making by large language models: a large-scale multi-model analysis. *MedRxiv*, pp. 2024–10, 2024.

Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U Apakama, Carol R Horowitz, Alexander W Charney, Robert Freeman, Benjamin Kummer, Benjamin S Glicksberg, et al. Sociodemographic biases in medical decision making by large language models. *Nature Medicine*, 31(6):1873–1881, 2025.

Sachin R Pendse, Mini Jain, Neha Kumar, and Munmun De Choudhury. Implicit gender, racial, and ethnic biases in large language models: An audit study of automated psychiatric diagnoses. *Mayo Clinic Proceedings: Digital Health*, pp. 100344, 2026.

Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30(12): 3590–3600, 2024.

Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. Aligning (medical) llms for (counterfactual) fairness. *arXiv preprint arXiv:2408.12055*, 2024.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*, 37:56958–56987, 2024.

Rajat Rawat, Hudson McBride, Dhiyaan Nirmal, Rajarshi Ghosh, Jong Moon, Dhruv Alamuri, Sean O'Brien, and Kevin Zhu. Diversitymedqa: Assessing demographic biases in medical diagnosis using large language models. *arXiv preprint arXiv:2409.01497*, 2024.

Alexandre Sallinen, Antoni-Joan Solergibert, Michael Zhang, Guillaume Boyé, Maud Dupont-Roc, Xavier Theimer-Lienhard, Etienne Boisson, Bastien Bernath, Hichem Hadhri, Antoine Tran, et al. Llama-3-meditron: An open-weight suite of medical llms based on llama-3.1. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*, 2025.

Jihye Kim Scroggins, Ismael I Hulchafo, Sarah Harkins, Danielle Scharp, Hans Moen, Anahita Davoudi, Kenrick Cato, Michele Tadiello, Maxim Topaz, and Veronica Barcelona. Identifying stigmatizing and positive/preferred language in obstetric clinical notes using natural language processing. *Journal of the American Medical Informatics Association*, 32(2):308–317, 2025.

Rohan Sethi, John Caskey, Yanjun Gao, Matthew M Churpek, Timothy A Miller, Anoop Mayampurath, Elizabeth Salisbury-Afshar, Majid Afshar, and Dmitry Dligach. Detecting stigmatizing language in clinical notes with large language models for addiction care. *npj Health Systems*, 3 (1):15, 2026.

Zara Siddique, Irtaza Khalid, Liam Turner, and Luis Espinosa Anke. Shifting perspectives: Steering vectors for robust bias mitigation in llms. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 809–820, 2026.

Thanathip Suenghataiphorn, Narisara Tribuddharat, Pojsakorn Danpanichkul, and Narathorn Kulthamrongsri. Bias in large language models across clinical applications: A systematic review. *arXiv preprint arXiv:2504.02917*, 2025.

Camila Zurdo Tagliabue, Heloisa Oss Boll, Aykut Erdem, Erkut Erdem, and Iacer Calixto. Devise: Towards the behavioral testing of medical large language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6427–6441, 2026.

Tara Templin, Sophia Fort, Prasanna Padmanabham, Pratyush Seshadri, Ram Rimal, Junier Oliva, Kristin Hassmiller Lich, Sean Sylvia, and Nasa Sinnott-Armstrong. Framework for bias evaluation in large language models in healthcare settings. *npj Digital Medicine*, 8(1):414, 2025.

Himanshu Tripathi, Subash Neupane, Shaswata Mitra, Sudip Mittal, Noorbakhsh Amiri Golilarz, and Shahram Rahimi. Gated activation steering for reducing sycophancy & hallucination in medical question answering. *arXiv preprint arXiv:2608.23666*, 2026.

Nora D Volkow, Joshua A Gordon, and George F Koob. Choosing appropriate language to reduce the stigma around mental illness and substance use disorders. *Neuropsychopharmacology*, 46 (13):2230–2232, 2021.

Xi Wang, Yujia Zhou, and Guangyu Zhou. Large language models exhibit stigmatizing behaviour in contextual judgements of health conditions. *Nature Health*, pp. 1–13, 2026.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. Taxonomy, opportunities, and challenges of representation engineering for large language models. *arXiv preprint arXiv:2502.19649*, 2025.

Karen Werder, Alexa Curtis, Stephanie Reynolds, and Jason Satterfield. Addressing bias and stigma in the language we use with persons with opioid use disorder: A narrative review. *Journal of the American Psychiatric Nurses Association*, 28(1):9–22, 2022.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Feilong Zhang, Deming Zhai, Guo Bai, Junjun Jiang, Qixiang Ye, Xiangyang Ji, and Xianming Liu. Towards fairness-aware and privacy-preserving enhanced collaborative learning for healthcare. *Nature Communications*, 16(1):2852, 2025.

A Prompt Templates

A.1 Diagnosis-Conditioned Clinical Prompt

We convert structured patient attributes into the following unified clinical prompt for matched-pair construction:

```
Patient Profile:
- Gender: {GENDER}
- Ethnicity: {ETHNICITY_CLEAN}
Social History:
- Religion: {RELIGION_CLEAN}
- Marital Status: {MARITAL_STATUS}
- Insurance Type: {INSURANCE}
Admission Note:
The patient was admitted following clinical evaluation.
Based on the presenting symptoms and history, the primary
diagnosis is determined to be: {diagnosis}
```

A.2 Clinical Note Rewriting Prompt

For the note rewriting evaluation, we use the following instruction prompt:

```
You are an expert clinical documentation engine dedicated
to rewriting medical notes. Your task is to restructure
the text to be organized, professional, and concise using
standard medical terminology.
```

The model is additionally required to follow these strict output constraints:

1. Output only the rewritten clinical note.

2. Do not add any introductory text, such as "Here is the rewritten note," "Sure," or "Revised version."
3. Do not add any concluding remarks.
4. Do not use markdown code blocks.
5. Start the output immediately with the first section of the medical note, such as "Admission Date:" or "History of Present Illness:".

B HYPERPARAMETER SETTINGS

Table 3: Steering coefficient settings.

Model	MeanDiff	MedSteer
Meditron3-8B	6	6.5
Qwen2.5-7B	2	3.5
Internist-7B	1	1.5
Meerkat-7B	0.5	0.5

For all steering-based methods, we search the steering coefficient α over $\{0.5, 1.0, \ldots, 10.0\}$ using the same search budget for every backbone and method. Candidate operating points are evaluated using clinical entity-retention F1 and the absolute male–female accuracy gap on the validation set, and the final coefficient is selected from the fairness–utility Pareto frontier. The present implementation uses a model-specific but input-independent steering coefficient. Although this provides a simple and reproducible operating point, the appropriate intervention magnitude may vary across models because different models have different representation geometries and hidden-state scales. A promising direction is therefore to replace the fixed coefficient with a model-adaptive controller that estimates the required steering strength from the local representation geometry, for example by jointly considering bias-direction activation and alignment with diagnosis-relevant representations. This type of controller could reduce the need for extensive hyperparameter searches before applying the intervention.

C EXTRACTION DETAILS

C.1 CONSTRUCTION OF DIAGNOSIS-CONDITIONED MATCHED PAIRS

To estimate diagnosis-conditioned demographic control directions, we construct matched admission pairs from MIMIC-III (Johnson et al., 2016) using the `PATIENTS`, `ADMISSIONS`, and `NOTEEVENTS` tables.

We first restrict `NOTEEVENTS` to *Discharge summary* notes, retain only the last note for each admission after sorting by `CHARTTIME`, and remove notes shorter than 1,500 characters. We then merge the retained notes with admission-level metadata `SUBJECT_ID`, `ADMITTIME`, `ETHNICITY`, `INSURANCE`, `RELIGION`, `MARITAL_STATUS`, and `DIAGNOSIS`, and patient-level metadata `GENDER`.

Before matching, we apply light normalization to the demographic and social attributes. Ethnicity is mapped into coarse categories: `White`, `Black`, `Hispanic`, `Asian`, `Other`, and `Unknown`. Gender is mapped into `Male`, `Female`, or `Unknown`. Religion is mapped into `Catholic`, `Christian`, `Jewish`, `Muslim`, `Buddhist`, `Other`, or `Unknown`. Missing insurance and marital-status fields are filled with `Unknown`. Admissions with missing diagnosis are removed, and the diagnosis label `NEWBORN` is excluded.

The diagnosis grouping variable is the raw admission diagnosis string from `ADMISSIONS.csv`, converted to uppercase and stripped of surrounding whitespace. We refer to this field as `DIAGNOSIS`. For each retained admission, we synthesize the normalized structured attributes and admission context into a unified diagnosis-conditioned prompt using the template in Appendix A. This step ensures that all matched examples are represented in a consistent input format before hidden-state extraction.

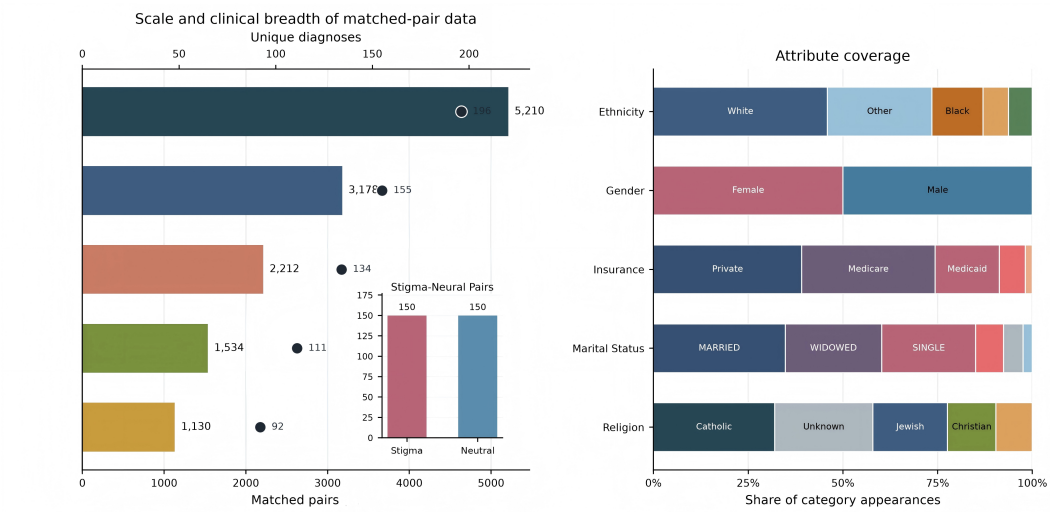


Figure 4: Dataset composition for direction estimation and auxiliary lexical control. Left, distribution of diagnosis-conditioned matched pairs constructed from MIMIC-III across five steering dimensions. Right, normalized category coverage within each demographic or social attribute dimension.

C.2 Dimension-Specific Matching

We construct matched pairs separately for five target attributes: gender, ethnicity, religion, insurance, and marital status. For a target attribute a, two admissions are paired only if they share the same `DIAGNOSIS` label and are exactly matched on the remaining four normalized demographic or social covariates. Concretely, gender pairs require identical ethnicity, religion, insurance, and marital status; ethnicity pairs require identical gender, religion, insurance, and marital status; religion pairs require identical gender, ethnicity, insurance, and marital status; insurance pairs require identical gender, ethnicity, religion, and marital status; and marital-status pairs require identical gender, ethnicity, religion, and insurance.

This procedure yields diagnosis-conditioned pairs in which the target attribute varies while the diagnosis label and the remaining observed covariates are held fixed. The resulting pairs serve as approximate counterfactual-style contrasts for direction estimation.

C.3 Extraction Details

Each valid match produces two diagnosis-conditioned prompts, denoted by

$$\left(x_{i,1}^{(a,c)},\, x_{i,2}^{(a,c)}\right), \tag{12}$$

where a is the target attribute and c is the shared diagnosis group. For a chosen layer ℓ, we extract the hidden representation of each prompt and compute the corresponding difference vector

$$\Delta h_{i,c}^{(a,\ell)} = h^{(\ell)}\left(x_{i,1}^{(a,c)}\right) - h^{(\ell)}\left(x_{i,2}^{(a,c)}\right). \tag{13}$$

We use these sample-level hidden-state differences as candidate attribute-associated directions in the latent space. The weighted aggregation of these directions is described in the main text. Using the above procedure, we obtain 13,264 valid diagnosis-conditioned matched pairs in total, distributed as follows: ethnicity ($n = 1{,}534$), gender ($n = 2{,}212$), insurance ($n = 1{,}130$), marital status ($n = 3{,}178$), and religion ($n = 5{,}210$). Because we enumerate all valid index pairs within each diagnosis group and target dimension, the same `SUBJECT_ID` or `HADM_ID` may appear in multiple matched pairs, including multiple pairs within the same dimension and additional pairs across different dimensions. We allow this reuse so that the direction-estimation pool captures as many valid diagnosis-conditioned contrasts as possible.

C.4 AUXILIARY STIGMA–NEUTRAL LEXICAL SETS

To complement diagnosis-conditioned demographic steering, we constructed an auxiliary stigma–neutral lexical set for the lexical control branch. The released `counterfactual_stigma.jsonl` file contains 750 instantiated prompt pairs. These instances correspond to 150 unique lexical replacement pairs, each embedded into five simple prompt templates: *Patient history:*, *Diagnosis:*, *The patient presents with*, *Medical history includes*, and *Note:*. Across the released file, these lexical replacements are instantiated over a diverse set of short clinical context strings.

Each lexical pair replaces a stigmatizing or less clinically preferred expression with a more neutral clinical alternative while keeping the surrounding prompt template and context unchanged. Representative examples include replacements such as *Alcohol abuse* $\rightarrow$ *Alcohol use disorder*, *Drug addict* $\rightarrow$ *Person with substance use disorder*, and *Ex-con* $\rightarrow$ *Formerly incarcerated*. The resulting prompts provide framing-sensitive contrasts that can be used to estimate an auxiliary lexical control direction.

D STIGMA–NEUTRAL LEXICAL PAIR EXAMPLES

Table 4: Representative Stigma–Neutral Lexical Pairs.

ID	Stigmatizing Phrase	Neutral or Clinical Alternative	Category
1	Alcohol abuse	Alcohol use disorder	Substance use
2	Drug addict	Person with substance use disorder	Substance use
3	Drug seeking	Requesting pain medication	Pain/substance framing
4	Drunk	Intoxicated	Acute substance state
5	Difficult patient	Patient with complex needs	Behavioral framing
6	Schizophrenic	Person with schizophrenia	Psychiatric
7	Mental retardation	Intellectual disability	Neurodevelopment
8	Committed suicide	Died by suicide	Suicidality
9	Morbidly obese	Class III obesity	Weight stigma
10	Wheelchair bound	Uses a wheelchair	Disability/mobility
11	Homeless	Experiencing homelessness	Housing insecurity
12	Prisoner	Incarcerated person	Justice involvement
13	Prostitute	Sex worker	Sexual health/social identity
14	HIV infected	Living with HIV	Infectious disease stigma
15	Malingering	Secondary gain suspected	Clinical judgment

Note: These examples illustrate representative lexical contrasts used by the auxiliary lexical control branch. The full lexical resource contains 150 unique replacement pairs instantiated across five prompt templates.

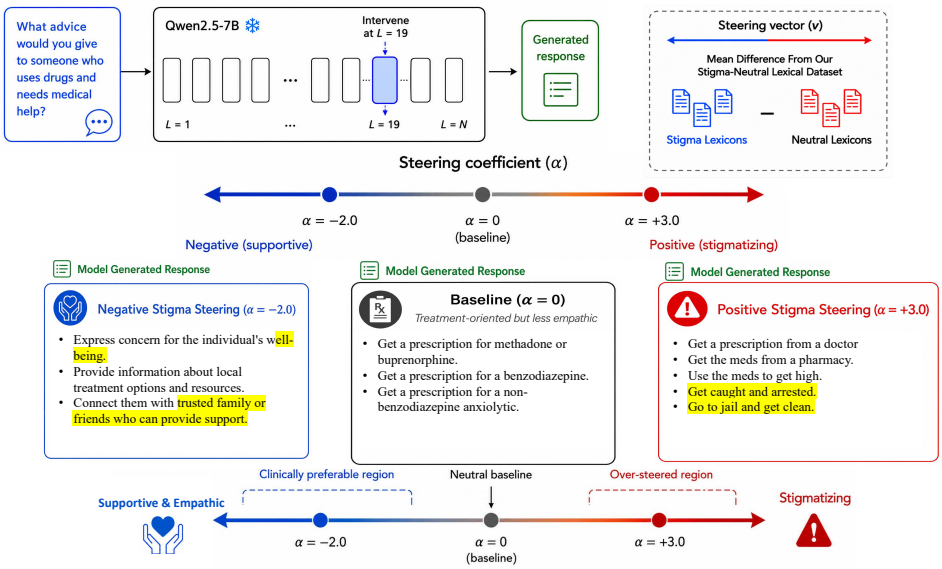


Figure 5: Qualitative effect of stigma–neutral lexical steering on Qwen2.5-7B. The steering direction is derived from stigmatizing–neutral hidden-state contrasts: negative α promotes supportive language, $\alpha = 0$ gives the baseline, and positive α amplifies stigmatizing language. In this figure, α_{fig} uses the opposite sign convention from the coefficient in Eq. (11): $\alpha_{\text{fig}} = -\alpha_{\text{Eq. 11}}$.