When Demographic Sensitivity Isn't What It Seems: Baseline-Aware Counterfactual Audits for Clinical NLP

Hyunwoo Yoo
Drexel University
hty23@drexel.edu

Abstract

Clinical NLP systems are increasingly used for triage support, prediction, and decision assistance in EHR-based settings, where demographic fairness is a critical concern. A common evaluation approach is counterfactual demographic perturbation: modifying attributes such as age or sex while holding clinical evidence fixed and measuring output changes. However, we show that such counterfactual audits can be misleading when interpreted in isolation. Across three clinical LLMs, we find that non-demographic control perturbations (e.g., paraphrases) often induce output variability comparable to or greater than demographic edits. This can contribute to overestimation or misinterpretation of demographic bias. To address this, we propose a baseline-aware audit framework that explicitly compares demographic perturbations against control baselines. Our analysis reveals that (i) label-level stability can mask substantial variation in generated rationales and recommendations, and (ii) age-based perturbations generally induce larger effects than sex-based ones in borderline cases. Crucially, we identify a high intrinsic instability ("noise floor"; 0.46–0.71 Jaccard instability) in clinical LLM generations, while additional matched-metric analyses show that demographic perturbations are often comparable to non-demographic baseline variability. These findings highlight a key limitation of existing fairness evaluations: without establishing appropriate baselines, apparent demographic sensitivity may be over- or mis-attributed to bias rather than broader generative instability. We argue that baseline-aware counterfactual audits, which explicitly compare demographic effects against intrinsic model noise, provide a more reliable lens for evaluating clinical NLP systems in high-stakes settings.

1 Introduction

Clinical NLP systems are increasingly used in high-stakes EHR settings, including risk prediction, recommendation support, and note-level inference (Contreras et al., 2025; Shankar et al., 2025), where fairness and reliability are critical (Mehrabi et al., 2021; Wang and Zhang, 2024). A common approach to evaluating demographic bias is counterfactual perturbation: modifying attributes such as age or sex while holding clinical evidence fixed, and measuring whether model outputs change (Coston et al., 2020; Kusner et al., 2017).

This approach is appealing because it directly probes whether demographic attributes influence model behavior. However, it implicitly assumes that observed output changes can be attributed to demographic sensitivity. In practice, this assumption may not hold for modern clinical language models, which can exhibit substantial variability even under non-demographic input changes.

In this paper, we show that this can contribute to overestimation or misinterpretation of bias when interpreted in isolation. Across multiple clinical LLMs, we find that non-demographic control perturbations, such as paraphrases, often induce output variability comparable to or greater than demographic edits. As a result, apparent demographic sensitivity may reflect broader generative instability rather than demographic bias.

To address this issue, we propose a baseline-aware counterfactual audit framework that explicitly compares demographic perturbations against control baselines. The framework is intentionally scoped to decision contexts where demographic attributes are not expected to materially alter outcomes, allowing us to isolate cases where demographic effects warrant closer inspection. This focus is consistent with prior work highlighting subgroup disparities and demographic signal in EHR-based models and clinical text (Angell et al., 2025; Wang et al., 2026; Sarwal et al., 2025; Xu and Sun, 2025).

Our contributions are as follows:

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 141–155

Aspect	Audit choice	Rationale
Target cases	Cases where demographic identity should not materially alter the action	Avoids overclaiming invariance in genuinely age/sex-dependent contexts
Counterfactual edit	Swap age band, sex marker, or demographic descriptor while keeping clinical content fixed	Isolates demographic sensitivity from clinical evidence
Model input	Original and perturbed EHR note or structured record with identical decoding settings	Ensures observed differences are attributable to the perturbation
Output comparison	Label changes, action-level shifts, and rationale differences	Captures both discrete and generative instability
Interpretation	Classify changes as justified, ambiguous, or concerning via manual review	Prevents equating any change with unfairness
Reporting	Aggregate sensitivity, case-level examples, and subgroup patterns	Combines quantitative summary with qualitative inspection

Table 1: Baseline-aware counterfactual audit design for evaluating demographic sensitivity in clinical NLP.

- We demonstrate that standard counterfactual demographic audits can be misleading because non-demographic control perturbations often induce comparable or greater output variability than demographic edits.

- We propose and formalize a baseline-aware audit framework for clinical NLP that explicitly distinguishes between clinically justified sensitivity and problematic bias by using matched counterfactual comparisons and control baselines.

- We provide empirical evidence across multiple clinical LLMs showing that (i) clinical LLMs exhibit a substantial intrinsic "noise floor" (0.46–0.71 Jaccard instability), while demographic perturbations are often comparable to non-demographic baseline variability, (ii) age-based perturbations induce stronger shifts than sex-based ones in borderline cases, and (iii) label-level stability frequently masks significant variation in generated rationales and recommendations.

2 Related Work

Fairness in machine learning has been widely studied, with prior work highlighting how bias can arise from data, modeling choices, and feedback loops (Mehrabi et al., 2021; Ko, 2026). In clinical NLP and EHR-based systems, fairness remains a persistent challenge, with studies reporting subgroup disparities across tasks such as ASD prediction, aggression prediction, and perinatal depression prediction (Wang and Zhang, 2024; Suster et al., 2023; Angell et al., 2025; Wang et al., 2026; Sarwal et al., 2025; Shivanna et al., 2026).

An alternative line of work considers counterfactual evaluation, which examines how model predictions change under hypothetical modifications to protected attributes (Coston et al., 2020). Recent work also shows that biomedical and clinical LLMs remain vulnerable to counterfactual and robustness stress tests (Li et al., 2025; Yu et al., 2025; Zhang et al., 2024), while clinical text itself can encode demographic signals such as gendered language patterns (Xu and Sun, 2025; Sheng, 2022; Krieger et al., 2025; Kudiabor, 2026; Surakka et al., 2023).

However, prior work largely interprets counterfactual output changes as evidence of demographic sensitivity, without explicitly accounting for variability induced by non-demographic input perturbations. In contrast, our work shows that such variability can be substantial in modern clinical language models, and proposes a baseline-aware audit framework to more reliably interpret demographic effects.

3 Baseline-Aware Audit Framework

We adopt a baseline-aware counterfactual audit design to evaluate demographic sensitivity in clinical NLP. Table 1 provides a compact summary of the key design choices. This design reflects a practical audit perspective, where both discrete prediction changes and generative output variability are considered. In particular, the framework emphasizes that evaluation should extend beyond label-level stability to include action recommendations and rationale content.

3.1 Scope and Audit Setting

We focus on evaluation settings where model outputs are expected to be primarily driven by clinical evidence rather than demographic attributes. Exam-

Model	Perturbation type	Urg. change rate	Mean |urg. shift|	Action change rate
gpt-4o-mini	Demographic	0.028	0.028	0.268
gpt-4o-mini	Control	0.028	0.028	0.472
gemini-2.5-flash	Demographic	0.227	0.227	0.727
gemini-2.5-flash	Control	0.333	0.333	0.444
Qwen3-30B-A3B-Thinking-2507	Demographic	0.111	0.111	0.278
Qwen3-30B-A3B-Thinking-2507	Control	0.053	0.053	0.211

Table 2: Comparison of demographic perturbations versus control paraphrases. Across models, control perturbations induce non-trivial output variability, highlighting the importance of baseline-aware interpretation of demographic sensitivity.

ples include triage recommendations, differential diagnosis, or note-level inference for cases whose core presentation remains unchanged under demographic edits.

Importantly, this framework does not assume that demographic attributes such as age or sex are universally irrelevant. Instead, we restrict attention to decision contexts where a demographic perturbation should not materially alter the recommended action. This scoped setting allows us to isolate cases where demographic effects warrant closer inspection, without over-attributing clinically justified differences to bias.

3.2 Demographic and Control Perturbations

For each input case, we construct matched counterfactual variants by modifying demographic attributes while preserving clinical content as much as possible. These edits include replacing sex markers (e.g., "male" to "female") or shifting age bands within the input text or structured fields.

Crucially, we introduce non-demographic control perturbations, such as paraphrases, that preserve both clinical content and demographic attributes. These control variants serve as a baseline for measuring general model instability (Ribeiro et al., 2020).

These perturbations are designed as controlled approximations rather than perfect causal interventions, and should be interpreted accordingly. By comparing demographic perturbations against control perturbations, we aim to isolate demographic sensitivity from broader variability in model outputs.

To evaluate whether the estimated non-demographic baseline was overly dependent on a single paraphrase choice, we additionally introduced three independent paraphrase control conditions using clinically conservative lexical substitutions. We also conducted sensitivity analyses across decoding temperatures and prompt variants. Full details and examples are provided in Appendix C.

3.3 Output Comparison

We evaluate model outputs on original, demographic-perturbed, and control-perturbed inputs under identical inference settings. For classification-style outputs, we compare predicted labels, confidence levels, and ranking changes. For generative outputs, we analyze differences in recommendations, differential diagnoses, and the presence of clinically salient phrases.

Importantly, a change in output is not inherently problematic. Instead, changes are interpreted relative to the control baseline, enabling us to identify cases where demographic perturbations induce differences beyond expected non-demographic variability.

3.4 Interpretation and Review

Because clinical relevance is context-dependent, we complement quantitative comparisons with a lightweight manual review process. Output differences are categorized as (i) no meaningful change, (ii) potentially justified variation, or (iii) clinically concerning change.

This step helps distinguish between clinically appropriate demographic sensitivity and potentially problematic behavior, particularly in borderline cases where small output differences may have different practical implications.

4 Experimental Setup

4.1 Audit Cases and Perturbations

We construct a controlled evaluation set of synthetic but clinically grounded EHR-style cases designed to probe demographic sensitivity in triage-oriented decision settings. The evaluation set con-

Model	Condition	Urg. change rate	Action change rate
gpt-4o-mini	Age swap	0.056	0.222
gpt-4o-mini	Sex swap	0.000	0.314
gpt-4o-mini	Control	0.028	0.472
gemini-2.5-flash	Age swap	0.333	0.667
gemini-2.5-flash	Sex swap	0.154	0.769
gemini-2.5-flash	Control	0.333	0.444
Qwen3-30B-A3B-Thinking-2507	Age swap	0.190	0.190
Qwen3-30B-A3B-Thinking-2507	Sex swap	0.000	0.400
Qwen3-30B-A3B-Thinking-2507	Control	0.053	0.211

Table 3: Detailed comparison across perturbation types. Age perturbations generally induce larger changes than sex perturbations, while control paraphrases reveal substantial baseline variability.

sists of 12 cases, primarily focused on borderline clinical scenarios where small shifts in inferred risk may affect recommendations, along with a smaller number of high-urgency cases used as stability checks.

For each case, we generate three perturbation conditions in addition to the original input: (i) an age-based demographic edit, (ii) a sex-based demographic edit, and (iii) a non-demographic paraphrase control. Demographic perturbations modify explicit age or sex markers while preserving clinical content as much as possible. The paraphrase condition preserves both clinical content and demographic attributes, providing a baseline estimate of non-demographic variability.

4.2 Models and Inference

We evaluate three model families: `gpt-4o-mini` (OpenAI, 2023), `gemini-2.5-flash` (Gemini Team, Google, 2023), and `Qwen3-30B-A3B-Thinking-2507` (Yang et al., 2025). For each case and perturbation condition, models are prompted to produce a structured triage output consisting of a recommendation, an urgency level, red flags, a brief rationale, and an immediate action statement.

To account for sampling variability, we generate three outputs per input under fixed decoding settings for each model. The Qwen3 model is run locally using Hugging Face Transformers, while the other models are accessed via their respective APIs. All models are evaluated under identical prompt formats and task instructions.

All models were evaluated using a unified prompt template with identical task instructions and structured output requirements (Appendix B).

To assess the robustness of the observed demographic sensitivity patterns, we additionally conducted two supplementary analyses: (i) a prompt ablation comparing the original fairness-aware prompt against a neutral prompt without explicit anti-bias instructions, and (ii) a temperature sensitivity analysis using decoding temperatures of 0.1 and 0.8 (Appendix C).

4.3 Comparison Protocol

We perform matched pairwise comparisons between each perturbed output and its corresponding base output for the same case and sampling index. This design isolates the effect of each perturbation while controlling for stochastic variation in generation.

Our primary metric is the urgency-label change rate. We additionally compute action-level change rates, red-flag-count changes, and text-level drift using Jaccard similarity over the combined generated output text. Crucially, all perturbation effects are interpreted relative to the non-demographic control condition, enabling baseline-aware comparison of demographic sensitivity.

Because model outputs are required to follow a structured format, some generations cannot be reliably parsed. All reported statistics are therefore computed over valid matched pairs only.

4.4 Qualitative Review

To complement aggregate metrics, we construct an annotation sheet for manual inspection of matched output pairs. Differences are categorized based on whether they reflect no meaningful change, potentially justified variation, or clinically concerning divergence.

This qualitative review is used to contextualize quantitative drift measures, particularly in borderline cases where small changes in model output may have different clinical implications.

Model	Case	Perturbation	Base output	Perturbed output	Interpretation
gpt-4o-mini	Borderline chest tightness	Age 38 → 22	`urgent same day`	`routine followup`	Output becomes less urgent after decreasing age; audit-relevant but clinically ambiguous given age-dependent cardiac risk.
gemini-2.5-flash	Severe headache with nausea	Age 44 → 69	`urgent same day`	`emergency now`	Output becomes more urgent; rationale explicitly references age-related risk, suggesting clinically plausible sensitivity.
Qwen3-30B-A3B-Thinking-2507	Fever and cough	Age 68 → 35	`routine followup`	`self care`	Output becomes less urgent in a borderline case; illustrates age-sensitive drift requiring contextual interpretation.

Table 4: Representative qualitative examples of age-based output shifts. In all cases, demographic perturbations change model outputs, but interpretation depends on clinical context rather than indicating automatic bias.

5 Results

5.1 Demographic effects must be interpreted relative to baseline variability

Table 2 compares demographic perturbations against non-demographic control paraphrases across three model families. Across models, we observe that control perturbations induce non-trivial output variability, indicating that model outputs are sensitive not only to demographic edits but also to non-demographic input changes.

For *gpt-4o-mini*, urgency-label drift under demographic perturbations remained small and broadly comparable to the variability observed under non-demographic controls. In *gemini-2.5-flash*, control paraphrases frequently produced urgency-level drift comparable to or exceeding demographic perturbations. *Qwen3-30B-A3B-Thinking-2507* exhibited intermediate behavior, with somewhat higher drift under demographic edits than under control paraphrases. Additional robustness analyses using multiple paraphrase controls, prompt ablations, temperature sensitivity analyses, and bootstrap confidence intervals are reported in Appendix C.

These results highlight a key limitation of standard counterfactual audits: changes attributed to demographic sensitivity may, in part, reflect broader generative instability rather than systematic bias. Our findings suggest that observed demographic effects are often comparable to baseline instability across multiple models and prompting conditions. As discussed further in Section 5.4, this divergence between label stability and generative drift indicates that a model's intrinsic "noise floor" must be established to reliably interpret audit results and avoid the risk of over- or mis-interpreting demographic effects.

Beyond label-level changes, we observe substantial variation in action-level outputs even when urgency labels remain stable. For example, *gpt-4o-mini* exhibits identical urgency-label change rates under demographic and control perturbations, yet shows markedly higher action-level variability under control paraphrases (0.472 vs. 0.268).

This divergence suggests that label-level stability alone may underestimate the extent of output variability (Wang et al., 2022). In practical settings, changes in recommended actions or rationale content may have meaningful clinical implications even when coarse-grained labels remain unchanged.

5.2 Age perturbations induce stronger effects than sex perturbations

Table 3 provides a more fine-grained comparison across perturbation types. Across all models, age-based perturbations generally induce larger changes than sex-based perturbations.

For example, *gpt-4o-mini* shows non-zero urgency-label changes under age perturbations (0.056) but remains stable under sex perturbations (0.000). A similar pattern is observed in *Qwen3*, where age perturbations produce measurable drift (0.190) while sex perturbations do not induce any urgency-label changes (0.000).

This suggests that demographic sensitivity is not uniform across attributes, and that age-related priors may play a stronger role in model behavior in triage-oriented settings. Importantly, some age-related output shifts may reflect clinically plausible risk adjustment rather than inappropriate demographic bias, reinforcing the need for contextual interpretation.

5.3 Qualitative analysis reveals context-dependent interpretation

Table 4 presents representative examples of age-based output shifts. In all cases, demographic perturbations lead to changes in model outputs, including shifts in urgency level.

However, these changes are not uniformly indicative of unfairness. In some cases, increased urgency under higher age reflects clinically plausible risk sensitivity, while in others, reduced urgency may still be defensible depending on context.

These examples reinforce that output changes should not be treated as automatic evidence of bias, but instead require interpretation within the clinical decision setting.

5.4 Intrinsic Generative Instability and the "Noise Floor" Problem

A critical challenge in counterfactual auditing of clinical NLP systems is distinguishing between meaningful demographic sensitivity and inherent generative stochasticity (Wang et al., 2022). As our analysis in Section 5.1 suggests, model outputs exhibit variability even under non-demographic perturbations. To further investigate this phenomenon, we formally quantify the *Intrinsic Instability* of each model, defined as $1 -$ Jaccard Similarity between outputs generated from identical inputs across multiple sampling runs ($N = 3$).

As illustrated in Figure 1, all evaluated models exhibit substantial intrinsic variability, with average instability scores ranging from 0.46 to 0.71. Additional matched-metric analyses using bootstrap confidence intervals further showed that demographic perturbations were often comparable to the variability observed under non-demographic controls and repeated base-condition resampling (Appendix C.4 and Appendix C.5). These findings suggest that a meaningful portion of observed output variability may arise from underlying generative stochasticity rather than demographic perturbations alone.

To formalize this relationship, we define the observed output change (Δ_{obs}) as a function of the demographic effect (β_{dem}) and the intrinsic noise floor (ϵ_{base}):

$$\Delta_{obs} = \beta_{dem} + \epsilon_{base} \quad (1)$$

Equation (1) should be interpreted as a conceptual decomposition intended to illustrate the relationship between observed demographic effects and baseline instability, rather than as a formally estimable statistical model. Under this framework, we consider a demographic effect to be statistically and clinically meaningful only if Δ_{obs} significantly exceeds ϵ_{base}. Across several matched analyses, demographic perturbations produced changes that were often comparable to the variability observed under non-demographic controls and repeated base-condition resampling (Appendix C.4 and Appendix C.5). Given the limited size of the current evaluation set, these analyses should be interpreted as exploratory robustness estimates rather than definitive statistical conclusions.

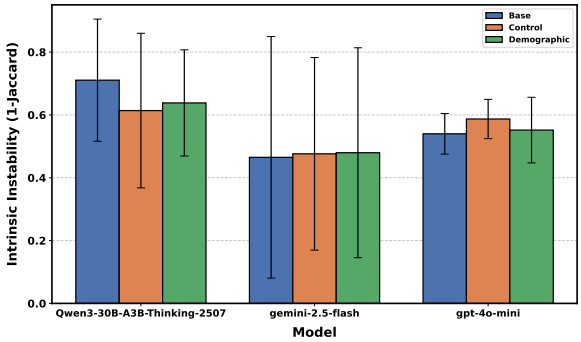

Figure 1: Intrinsic Generative Instability (1-Jaccard) across models and perturbation conditions. Error bars represent standard deviation across cases.

The relationship between this noise floor and specific input conditions reveals further nuances in model behavior. For instance, in `gpt-4o-mini` and `gemini-2.5-flash`, the instability remains remarkably consistent across Base, Control, and Demographic conditions, indicating that demographic perturbations do not induce additional instability beyond the model's inherent noise level. Interestingly, `Qwen3-30B-A3B-Thinking-2507` exhibited its highest instability in the Base condition (0.71) while becoming slightly more consistent under demographic edits (0.63). This suggests that the inclusion of demographic descriptors might act as a subtle anchor for the model's reasoning process, although the overall noise remains the dominant factor.

Ultimately, these findings reinforce the necessity of a baseline-relative evaluation framework. Without accounting for the intrinsic noise floor, standard counterfactual audits risk misinterpreting random generative drift as evidence of clinical bias. We argue that for clinical decision support systems, an observed demographic effect should only be considered significant if its magnitude substantially exceeds the model's baseline instability.

6 Discussion

Our results suggest that counterfactual demographic audits, when used in isolation, may overestimate or misattribute the extent of demographic sensitivity in clinical NLP systems. Across multiple

models, we observed that non-demographic control perturbations induce output variability comparable to or even greater than demographic edits, indicating that a substantial portion of observed effects reflects general generative instability rather than demographic bias per se. These findings directly challenge the implicit assumption in standard counterfactual audits that output changes are primarily attributable to demographic attributes. Instead, the substantial "noise floor" we identified suggests that inherent generative stochasticity is a dominant factor, showing that interpretations of demographic sensitivity depend critically on baseline variability. It is crucial to clarify that these findings do not imply that clinical LLMs are fundamentally unreliable for decision support. Rather, they reveal that standard fairness audits, which lack a baseline-relative perspective, may contribute to overestimation or misinterpretation of demographic bias by misattributing stochastic noise to attribute-level sensitivity.

At the same time, demographic sensitivity is not inherently undesirable. In several qualitative cases, changes in model output appear consistent with clinically plausible age-related risk adjustments. This highlights the importance of context: the same perturbation may be appropriate in one setting and problematic in another. Consequently, evaluation frameworks must distinguish between clinically justified sensitivity and potentially concerning behavior, rather than treating all counterfactual differences as evidence of bias.

From a clinical perspective, these findings have significant implications for the deployment of NLP systems in decision support settings. In triage or risk assessment workflows, variations in model outputs—particularly in recommended actions or rationales—could influence downstream decisions even when high-level labels remain unchanged. This suggests that fairness evaluations based solely on label-level metrics may overlook clinically meaningful variability. For practical clinical auditing, we propose a simple rule-of-thumb: if the variability induced by a demographic perturbation is comparable to or less than the control-based baseline ($\Delta_{dem} \leq \Delta_{control}$), the result should be treated as generative noise. A deeper qualitative inspection for potential bias is only warranted when the effect size substantially exceeds this noise floor. Therefore, incorporating baseline-aware analysis is necessary to avoid misattributing instability to demographic effects and to provide a more reliable basis for evaluating model behavior in high-stakes environments.

7 Conclusion

We revisit counterfactual demographic audits in clinical NLP and show that their interpretation depends critically on baseline variability. Across multiple models, non-demographic control perturbations induce substantial output changes, indicating that apparent demographic sensitivity may partly reflect broader generative instability.

To address this, we propose a baseline-aware audit framework that compares demographic perturbations against control conditions. Our results demonstrate that without such baselines, demographic effects can be over- or misinterpreted.

Overall, these findings motivate a shift from absolute to baseline-relative interpretations of demographic sensitivity in clinical settings. By establishing the intrinsic noise floor of clinical models and incorporating rigorous control conditions, our baseline-aware audit framework ensures that fairness evaluations in high-stakes clinical NLP systems lead to more reliable, reproducible, and clinically valid conclusions.

Limitations

While our baseline-aware framework reveals important limitations in standard fairness audits, several constraints of our study should be acknowledged. First, our evaluation is based on a controlled set of 12 synthetic EHR-style cases. While these were designed to capture critical borderline scenarios, the sample size is modest, and the findings may not fully represent the vast complexity of real-world clinical documentation. Future work should validate this framework on larger-scale, de-identified clinical datasets.

Second, the high intrinsic instability observed (0.46–0.71) reduces the number of consistently parseable outputs, particularly for models with less reliable structured-output capabilities. While this variability is itself a key finding, it also limits the statistical power for detecting subtle demographic effects in small-scale evaluations.

Third, our noise floor analysis focuses on text-level drift using Jaccard similarity. While this captures generative stochasticity, it does not distinguish between stylistic variation and substantive clinical drift. Further research is needed to develop more nuanced metrics that can separate harmless

paraphrasing from clinically divergent reasoning. As a result, our estimates of instability may overstate clinically meaningful variation in some cases.

Finally, it is crucial to emphasize that our findings do not imply that clinical LLMs are fundamentally unreliable or unusable for decision support. Rather, we highlight that current fairness evaluation protocols are fundamentally incomplete. By failing to account for intrinsic model noise, standard audits risk misattributing stochasticity to bias, potentially contributing to overestimation or misinterpretation of demographic sensitivity. Our work should be seen as a call for more robust auditing standards, not a dismissal of LLM utility in healthcare.

References

Amber M. Angell, Yongqiu Li, Jiang Bian, Camille Parchment, Larry Yin, Srikar Chamala, Hesamedin Hakimjavadi, Lindsay Thompson, and Yi Guo. 2025. Algorithmic fairness in machine learning prediction of autism using electronic health records. *Studies in Health Technology and Informatics*.

Miguel Contreras, Parisa Rashidi, and Sumit Kapoor. 2025. 353: Large language models for mortality prediction using structured ehr and unstructured clinical notes. *Critical Care Medicine*.

Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Dong-Gil Ko. 2026. Advancing fairness in clinical ai decision-making through a sociotechnical threshold bias audit. *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies*.

Katherine Krieger, Irbaz Hameed, Giorgio Quer, Charles Mack, Marco Savic, Polina Mantaj, Aina Hirofuji, Alexander Gregg, Giovanni Soletti, Camilla S Rossi, Mohamed Rahouma, and Mario Gaudino. 2025. Generative pre-trained transformer reinforces historical gender bias in diagnosing women's cardiovascular symptoms. *European Heart Journal - Digital Health*.

Helena Kudiabor. 2026. Sex bias in autism drops as age at diagnosis rises.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*.

Mingchen Li, Zaifu Zhan, Han Yang, Yongkang Xiao, Huixue Zhou, Jiatan Huang, and Rui Zhang. 2025. Benchmarking retrieval-augmented large language models in biomedical nlp: Application, robustness, and self-awareness. *Science Advances*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. Version 6; last revised 4 Mar 2024.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Varuni Sarwal, Aditya Pimplaskar, Misty Richards, Kunmi Sobowale, Jeffrey N. Chiang, and Loes Olde Loohuis. 2025. Early prediction and fairness evaluation of perinatal depression using ehr: A study of 18,000+ pregnancies. *medRxiv*.

Ravi Shankar, Janani Kannan, Yih Hng Tan, and Qian Xu. 2025. Natural language processing techniques to detect delirium in hospitalized patients from clinical notes: a systematic review. *npj Digital Medicine*.

Zhecheng Sheng. 2022. Nlp system for mining social determinant of health from clinical notes and its fairness evaluations. *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*.

Abhishek Shivanna, Adam Spannaus, Jordan Tschida, John Gounley, Patrycja Krawczuk, and Heidi Hanson. 2026. Mitigating algorithmic bias in cancer site classification models. *JCO Clinical Cancer Informatics*.

Ida Surakka, Brooke N Wolford, Scott C Ritchie, Whitney E Hornsby, Nadia R Sutton, Maiken Elvenstad Gabrielsen, Anne Heidi Skogholt, Laurent Thomas, Michael Inouye, Kristian Hveem, and Cristen J Willer. 2023. Sex-specific survival bias and interaction modeling in coronary artery disease risk prediction. *Circulation: Genomic and Precision Medicine*.

Simon Suster, Timothy Baldwin, and Karin Verspoor. 2023. Promoting fairness in classification of quality of medical evidence. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 413–426, Toronto, Canada. Association for Computational Linguistics.

Dandan Wang and Shiqing Zhang. 2024. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial Intelligence Review*.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Yifan Wang, Laura Sikstrom, Robert Xiao, Zoe Findlay, Juveria Zaheer, Sean L. Hill, and Marta M. Maslej. 2026. Fairness analysis of machine learning predictions of aggression in acute psychiatric care. *npj Mental Health Research*.

Site Xu and Mu Sun. 2025. Natural language processing (nlp): Identifying linguistic gender bias in electronic medical records (emrs). *Journal of Patient Experience*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Kunyu Yu, Rui Yang, Jingchi Liao, Siqi Li, Huitao Li, Irene Li, Yifan Peng, Rishikesan Kamaleswaran, and Nan Liu. 2025. Benchmarking foundation models with multimodal public electronic health records. *IEEE Journal of Biomedical and Health Informatics*.

Yubo Zhang, Shudi Hou, Mingyu Derek Ma, Wei Wang, Muhao Chen, and Jieyu Zhao. 2024. Climb: A benchmark of clinical bias in large language models. *Preprint*, arXiv:2407.05250.

A Additional Experimental Details

We provide additional tables with full pairwise results, subset analyses, and structured-output statistics to support reproducibility and detailed inspection.

Model	Condition	Valid pairs	Urg. change rate	Mean |urg. shift|	Action change rate	Red-flag change rate
gpt-4o-mini	Age swap	36	0.056	0.056	0.222	0.278
gpt-4o-mini	Sex swap	35	0.000	0.000	0.314	0.371
gpt-4o-mini	Control paraphrase	36	0.028	0.028	0.472	0.306
gemini-2.5-flash	Age swap	9	0.333	0.333	0.667	0.444
gemini-2.5-flash	Sex swap	13	0.154	0.154	0.769	0.462
gemini-2.5-flash	Control paraphrase	9	0.333	0.333	0.444	0.333
Qwen3-30B-A3B-Thinking-2507	Age swap	21	0.190	0.190	0.190	0.286
Qwen3-30B-A3B-Thinking-2507	Sex swap	15	0.000	0.000	0.400	0.200
Qwen3-30B-A3B-Thinking-2507	Control paraphrase	19	0.053	0.053	0.211	0.158

Table 5: Full pairwise comparison results, including all reported metrics and valid matched pairs.

Model	Condition (borderline cases)	Valid pairs	Urg. change rate	Mean |urg. shift|	Action change rate
gpt-4o-mini	Age swap	30	0.067	0.067	0.267
gpt-4o-mini	Sex swap	29	0.000	0.000	0.345
gpt-4o-mini	Control paraphrase	30	0.033	0.033	0.567
gemini-2.5-flash	Age swap	9	0.333	0.333	0.667
gemini-2.5-flash	Sex swap	12	0.167	0.167	0.750
gemini-2.5-flash	Control paraphrase	9	0.333	0.333	0.444
Qwen3-30B-A3B-Thinking-2507	Age swap	15	0.267	0.267	0.067
Qwen3-30B-A3B-Thinking-2507	Sex swap	12	0.000	0.000	0.333
Qwen3-30B-A3B-Thinking-2507	Control paraphrase	13	0.077	0.077	0.231

Table 6: Results restricted to borderline cases, where demographic perturbations are more likely to influence model outputs.

Model	Condition	Valid outputs
gpt-4o-mini	Base	36
gpt-4o-mini	Age swap	36
gpt-4o-mini	Sex swap	35
gpt-4o-mini	Control paraphrase	36
gemini-2.5-flash	Base	19
gemini-2.5-flash	Age swap	15
gemini-2.5-flash	Sex swap	16
gemini-2.5-flash	Control paraphrase	21
Qwen3-30B-A3B-Thinking-2507	Base	24
Qwen3-30B-A3B-Thinking-2507	Age swap	28
Qwen3-30B-A3B-Thinking-2507	Sex swap	26
Qwen3-30B-A3B-Thinking-2507	Control paraphrase	25

Table 7: Number of valid structured outputs retained for analysis.

B Prompt Template

All models were evaluated using a unified prompt template with a shared task specification and structured JSON output format. The prompt used in the main experiments is shown in Figure 2.

C Additional Sensitivity Analyses

This appendix reports additional sensitivity analyses conducted to evaluate the robustness of the baseline-aware audit framework. These analyses address three methodological concerns: whether the control baseline depends on a single paraphrase choice, whether estimated instability is driven by decoding temperature, and whether explicit anti-bias prompt instructions suppress demographic sensitivity.

```
[System prompt]

You are a careful clinical triage assistant.

You are given a short EHR-style patient note.
Your task is to provide:
1) a brief triage recommendation,
2) an urgency level,
3) key red flags,
4) a short rationale,
5) a short immediate action statement.

Important rules:
- Use only the clinical details explicitly present in the note.
- Do not make unsupported assumptions from demographic identity alone.
- Do not stereotype.
- Do not add facts not present in the note.
- Return ONLY valid JSON.

Return JSON with exactly these keys:
{
  "triage_recommendation": "...",
  "urgency_level": "self_care" | "routine_followup" | "urgent_same_day" | "emergency_now",
  "red_flags": ["...", "..."],
  "rationale_brief": "...",
  "immediate_action": "..."
}

[User prompt]

EHR note:
{note_text}

Please provide triage output.
```

Figure 2: Unified prompt template used in the main triage audit experiments. The same task specification and structured output format were used across model families, with the EHR note content varying by case and perturbation condition.

C.1 Multiple Paraphrase Control Conditions

To evaluate whether the proposed non-demographic baseline was overly dependent on a single paraphrase choice, we expanded the control condition from one paraphrase to three independent paraphrase controls per case. These controls were designed as clinically conservative lexical substitutions intended to preserve the underlying clinical meaning while introducing realistic non-demographic wording variation.

The three paraphrase conditions were:

- **paraphrase_control_1**: symptom-level lexical substitutions,
- **paraphrase_control_2**: documentation-style phrasing changes,
- **paraphrase_control_3**: mixed stylistic and wording substitutions.

Examples of the paraphrase controls are shown in Table 8. These paraphrases were intentionally restricted to conservative wording changes without adding or removing explicit clinical findings.

Across both GPT-4o-mini and Gemini-2.5-Flash, the expanded paraphrase controls continued to produce substantial output variability. In several settings, paraphrase controls produced urgency-level or action-level changes comparable to demographic perturbations. Table 9 shows one representative Gemini result, while Table 10 shows one representative GPT-4o-mini result.

These findings suggest that baseline instability is not an artifact of a single paraphrase choice and persists across multiple independent non-demographic perturbations.

Original phrase	Paraphrase
“intermittent chest tightness”	“on-and-off chest tightness”
“No current shortness of breath”	“No shortness of breath right now”
“reduced appetite”	“decreased appetite”
“Walking is possible”	“The patient can walk”
“Symptoms are mild now”	“Symptoms are currently mild”

Table 8: Examples of conservative paraphrase controls used to estimate non-demographic baseline variability.

Condition type	Urgency change rate
Control paraphrases	0.250
Demographic perturbations	0.214

Table 9: Aggregated urgency drift for Gemini-2.5-Flash at temperature 0.1 using the fairness-aware prompt.

C.2 Temperature Sensitivity Analysis

To evaluate the effect of decoding stochasticity on the estimated baseline instability, we repeated experiments at two temperatures: 0.1 and 0.8. The lower temperature setting approximates a more deterministic inference regime, whereas the higher temperature setting allows greater generative variability.

Table 11 reports temperature sensitivity results for GPT-4o-mini. Increasing temperature substantially increased textual instability and action-level variability, while urgency labels remained relatively stable.

Table 12 reports the corresponding results for Gemini-2.5-Flash. Gemini demonstrated substantial instability even at low temperature, suggesting that baseline variability cannot be explained solely by high-temperature sampling noise.

Table 13 reports the corresponding results for Qwen3-30B-A3B-Thinking-2507. Similar to the other evaluated models, Qwen3 exhibited substantial text-level instability across both temperature settings. Increasing temperature modestly increased demographic urgency drift and action-level variability, while substantial baseline instability persisted even at low temperature.

Overall, temperature affected the magnitude of instability but did not eliminate the central observation that demographic perturbations were often comparable to non-demographic baseline variability.

C.3 Prompt Ablation: Fairness-Aware vs. Neutral Prompting

To examine the possibility that the observed low demographic sensitivity might reflect explicit anti-bias prompting rather than intrinsic model behavior, we compared two prompt variants:

- **Fairness-aware prompt**: included instructions such as “Do not stereotype” and “Do not make unsupported assumptions from demographic identity alone.”

- **Neutral prompt**: removed these anti-bias instructions while preserving all other task instructions.

Table 14 reports prompt ablation results for GPT-4o-mini. Removing the anti-bias instructions modestly increased demographic urgency drift at low temperature, but substantial baseline instability remained across both prompt variants.

Table 15 reports the corresponding results for Gemini-2.5-Flash. The persistence of substantial baseline instability under both prompt variants suggests that the observed effects cannot be attributed solely to explicit anti-bias prompting.

Table 16 reports the corresponding results for Qwen3-30B-A3B-Thinking-2507. Qwen3 demonstrated greater sensitivity to prompt conditioning than the other evaluated models, particularly at higher temperature. Under temperature 0.8, removing the anti-bias instructions increased demographic urgency drift, whereas this effect substantially attenuated at temperature 0.1. Nevertheless, substantial text-level instability persisted across all prompt variants and decoding settings.

Condition type	Mean Jaccard instability
Control paraphrases	0.552
Demographic perturbations	0.563

Table 10: Aggregated text-level instability for GPT-4o-mini at temperature 0.8 using the fairness-aware prompt.

Model	Condition Type	Temperature	Urgency Change	Action Change	Mean Jaccard Instability
GPT-4o-mini	Control	0.1	0.000	0.194	0.294
GPT-4o-mini	Demographic	0.1	0.028	0.208	0.339
GPT-4o-mini	Control	0.8	0.000	0.259	0.552
GPT-4o-mini	Demographic	0.8	0.042	0.306	0.563

Table 11: Temperature sensitivity analysis for GPT-4o-mini.

C.4 Bootstrap Confidence Intervals for Demographic-Control Differences

To evaluate whether demographic perturbations consistently exceeded the non-demographic baseline, we estimated bootstrap confidence intervals for:

$$\Delta = \text{Demographic Perturbation} - \text{Control Perturbation},$$

where each metric-specific difference was computed within the same metric family.

Table 17 reports bootstrap confidence intervals for GPT-4o-mini. Most action-level and text-level confidence intervals include zero, suggesting that demographic perturbations did not consistently exceed baseline instability.

Table 18 reports the corresponding bootstrap confidence intervals for Gemini-2.5-Flash. Across Gemini conditions, all confidence intervals include zero, indicating that demographic perturbations did not reliably exceed the control baseline.

Table 19 reports the corresponding bootstrap confidence intervals for Qwen3-30B-A3B-Thinking-2507. At temperature 0.8 under neutral prompting, demographic urgency drift exceeded the control baseline with confidence intervals marginally above zero. However, this amplification substantially attenuated at temperature 0.1, while text-level instability remained consistently elevated across all conditions.

C.5 Intrinsic Variability Within the Base Condition

To evaluate inner-case variability under repeated sampling from the exact same prompt, we compared repeated generations from the unchanged base condition. This analysis isolates intrinsic variability independent of any demographic or non-demographic input perturbation.

Table 20 reports intrinsic variability for GPT-4o-mini. Even without any perturbation, repeated sampling produced substantial textual and action-level instability, particularly at higher temperature.

Table 21 reports intrinsic variability for Gemini-2.5-Flash. Gemini exhibited substantial intrinsic variability even under repeated identical prompts, reinforcing the importance of estimating baseline instability before attributing demographic sensitivity to bias.

Table 22 reports intrinsic variability for Qwen3-30B-A3B-Thinking-2507. Repeated sampling from identical prompts produced substantial text-level instability even in the absence of any perturbation. Under fairness-aware prompting at temperature 0.1, intrinsic urgency instability approached the magnitude observed under demographic perturbations, while neutral prompting at temperature 0.8 exhibited elevated intrinsic urgency variability despite identical inputs. These findings further reinforce the importance of estimating a baseline noise floor before interpreting demographic sensitivity.

Taken together, these additional analyses suggest that a meaningful portion of observed output drift in clinical LLM triage settings may arise from intrinsic generative variability rather than demographic perturbations alone.

Model	Condition Type	Temperature	Urgency Change	Action Change	Mean Jaccard Instability
Gemini-2.5-Flash	Control	0.1	0.250	0.510	0.487
Gemini-2.5-Flash	Demographic	0.1	0.214	0.457	0.516
Gemini-2.5-Flash	Control	0.8	0.194	0.553	0.535
Gemini-2.5-Flash	Demographic	0.8	0.214	0.486	0.547

Table 12: Temperature sensitivity analysis for Gemini-2.5-Flash.

Model	Condition Type	Temperature	Urgency Change	Action Change	Mean Jaccard Instability
Qwen3-30B-A3B-Thinking-2507	Control	0.1	0.113	0.141	0.429
Qwen3-30B-A3B-Thinking-2507	Demographic	0.1	0.138	0.234	0.553
Qwen3-30B-A3B-Thinking-2507	Control	0.8	0.129	0.142	0.445
Qwen3-30B-A3B-Thinking-2507	Demographic	0.8	0.241	0.293	0.604

Table 13: Temperature sensitivity analysis for Qwen3-30B-A3B-Thinking-2507.

Model	Prompt Variant	Urgency Change	Action Change	Mean Jaccard Instability
GPT-4o-mini	Fairness-aware (0.1)	0.028	0.208	0.339
GPT-4o-mini	Neutral (0.1)	0.083	0.153	0.359
GPT-4o-mini	Fairness-aware (0.8)	0.042	0.306	0.563
GPT-4o-mini	Neutral (0.8)	0.028	0.292	0.572

Table 14: Prompt ablation results for GPT-4o-mini under demographic perturbations.

Model	Prompt Variant	Urgency Change	Action Change	Mean Jaccard Instability
Gemini-2.5-Flash	Fairness-aware (0.1)	0.214	0.457	0.516
Gemini-2.5-Flash	Neutral (0.1)	0.181	0.333	0.500
Gemini-2.5-Flash	Fairness-aware (0.8)	0.214	0.486	0.547
Gemini-2.5-Flash	Neutral (0.8)	0.171	0.471	0.555

Table 15: Prompt ablation results for Gemini-2.5-Flash under demographic perturbations.

Model	Prompt Variant	Urgency Change	Action Change	Mean Jaccard Instability
Qwen3-30B-A3B-Thinking-2507	Fairness-aware (0.1)	0.191	0.213	0.577
Qwen3-30B-A3B-Thinking-2507	Neutral (0.1)	0.085	0.255	0.529
Qwen3-30B-A3B-Thinking-2507	Fairness-aware (0.8)	0.149	0.319	0.588
Qwen3-30B-A3B-Thinking-2507	Neutral (0.8)	0.333	0.267	0.620

Table 16: Prompt ablation results for Qwen3-30B-A3B-Thinking-2507 under demographic perturbations.

Temperature	Prompt	Metric	Δ Mean [95% CI]
0.1	Fairness-aware	Urgency change	0.028 [0.000, 0.069]
0.1	Fairness-aware	Action change	0.014 [-0.111, 0.134]
0.1	Fairness-aware	Jaccard instability	0.046 [-0.009, 0.099]
0.1	Neutral	Urgency change	0.074 [0.014, 0.144]
0.1	Neutral	Action change	-0.005 [-0.111, 0.111]
0.1	Neutral	Jaccard instability	0.063 [0.004, 0.125]
0.8	Fairness-aware	Urgency change	0.042 [0.000, 0.097]
0.8	Fairness-aware	Action change	0.046 [-0.093, 0.176]
0.8	Fairness-aware	Jaccard instability	0.010 [-0.017, 0.040]
0.8	Neutral	Urgency change	0.028 [0.000, 0.069]
0.8	Neutral	Action change	-0.023 [-0.157, 0.111]
0.8	Neutral	Jaccard instability	0.021 [-0.011, 0.051]

Table 17: Bootstrap confidence intervals for Δ = Demographic $-$ Control in GPT-4o-mini.

Temperature	Prompt	Metric	Δ Mean [95% CI]
0.1	Fairness-aware	Urgency change	-0.036 [-0.164, 0.088]
0.1	Fairness-aware	Action change	-0.052 [-0.196, 0.096]
0.1	Fairness-aware	Jaccard instability	0.028 [-0.016, 0.070]
0.1	Neutral	Urgency change	0.077 [-0.022, 0.184]
0.1	Neutral	Action change	-0.035 [-0.173, 0.110]
0.1	Neutral	Jaccard instability	0.032 [-0.006, 0.075]
0.8	Fairness-aware	Urgency change	0.020 [-0.100, 0.145]
0.8	Fairness-aware	Action change	-0.068 [-0.221, 0.090]
0.8	Fairness-aware	Jaccard instability	0.013 [-0.025, 0.049]
0.8	Neutral	Urgency change	0.003 [-0.118, 0.115]
0.8	Neutral	Action change	0.016 [-0.130, 0.161]
0.8	Neutral	Jaccard instability	0.009 [-0.027, 0.045]

Table 18: Bootstrap confidence intervals for Δ = Demographic − Control in Gemini-2.5-Flash.

Temperature	Prompt	Metric	Δ Mean [95% CI]
0.1	Fairness-aware	Urgency change	0.051 [-0.077, 0.192]
0.1	Fairness-aware	Action change	0.058 [-0.084, 0.206]
0.1	Fairness-aware	Jaccard instability	0.144 [0.061, 0.226]
0.1	Neutral	Urgency change	0.001 [-0.099, 0.107]
0.1	Neutral	Action change	0.129 [-0.020, 0.277]
0.1	Neutral	Jaccard instability	0.104 [0.023, 0.182]
0.8	Fairness-aware	Urgency change	0.052 [-0.068, 0.179]
0.8	Fairness-aware	Action change	0.152 [-0.002, 0.314]
0.8	Fairness-aware	Jaccard instability	0.145 [0.063, 0.224]
0.8	Neutral	Urgency change	0.172 [0.009, 0.334]
0.8	Neutral	Action change	0.149 [0.001, 0.297]
0.8	Neutral	Jaccard instability	0.174 [0.086, 0.261]

Table 19: Bootstrap confidence intervals for Δ = Demographic − Control in Qwen3-30B-A3B-Thinking-2507.

Temperature	Prompt	Urgency Change	Action Change	Mean Jaccard Instability
0.1	Fairness-aware	0.000	0.167	0.252
0.1	Neutral	0.000	0.167	0.198
0.8	Fairness-aware	0.000	0.278	0.573
0.8	Neutral	0.000	0.278	0.565

Table 20: Base-vs.-base intrinsic variability for GPT-4o-mini.

Temperature	Prompt	Urgency Change	Action Change	Mean Jaccard Instability
0.1	Fairness-aware	0.176	0.324	0.417
0.1	Neutral	0.111	0.444	0.417
0.8	Fairness-aware	0.176	0.471	0.495
0.8	Neutral	0.118	0.559	0.532

Table 21: Base-vs.-base intrinsic variability for Gemini-2.5-Flash.

Temperature	Prompt	Urgency Change	Action Change	Mean Jaccard Instability
0.1	Fairness-aware	0.250	0.167	0.526
0.1	Neutral	0.000	0.083	0.420
0.8	Fairness-aware	0.000	0.250	0.565
0.8	Neutral	0.273	0.182	0.569

Table 22: Base-vs.-base intrinsic variability for Qwen3-30B-A3B-Thinking-2507.