Counterfactual Auditing of Cross-Cultural Variation in LLM-Generated Medical Advice

Hyunwoo Yoo **Gail Rosen**

Drexel University
{hty23, glr26}@drexel.edu

Abstract

Large language models (LLMs) are increasingly explored for patient-facing medical advice and symptom triage, yet their responses may shift when identical clinical evidence is paired with culturally marked patient descriptors. We present a counterfactual audit framework for evaluating cross-cultural variation in LLM-generated medical advice by isolating identity-related cues while holding clinical evidence constant. Our evaluation uses matched clinical vignettes, cross-regional and culturally marked prompt variants, repeated sampling, and structured comparison of urgency framing, safety recommendations, empathy, and escalation advice. Across multiple commercial and open-weight LLMs, we observe measurable identity-conditioned variation in both triage decisions and interactional framing. In several cases, culturally marked descriptors shift urgency assessments or escalation recommendations despite unchanged clinical evidence. While the magnitude and direction of these effects differ across models, the results suggest that LLM-generated medical advice remains sensitive to culturally linked identity cues in ways that may affect safety-critical guidance. Our results demonstrate how culturally grounded counterfactual auditing can help identify clinically unsupported variation while distinguishing potentially harmful shifts from appropriate communication adaptation in patient-facing medical advice.

1 Introduction

Large language models (LLMs) are increasingly explored for patient-facing medical advice and symptom triage, but safety-critical settings require more than factual correctness alone. In clinical guidance contexts, small differences in wording can affect perceived urgency or willingness to seek care, particularly when identical symptom descriptions are paired with culturally marked patient descriptors. Prior work on medical LLMs has highlighted harm, bias, and scientific grounding as distinct evaluation concerns (Singhal et al., 2023; Tam et al., 2024; Wang and Zhang, 2024). Related counterfactual audits further show that when demographic cues are varied while clinical evidence is held constant, models can produce different diagnoses, recommendations, or interactional framing (Omiye et al., 2023; Yang et al., 2024; Hanna et al., 2025).

This paper presents a counterfactual audit framework for analyzing cross-cultural variation in LLM-generated medical advice. Our focus is not to argue for culturally blind healthcare interactions: cultural competence, language access, and communication style can all matter in real clinical settings. Instead, we ask whether *matched* symptom scenarios, in which core clinical evidence remains fixed, lead to unjustified changes when prompts vary only in culturally marked self-identification, language, or locale. We study whether these cues shift urgency assessments, safety framing, empathy, or escalation advice in ways that are not supported by the underlying clinical facts.

We consider patient-facing medical advice and symptom-triage generation. Given a fixed vignette describing symptoms and context, a model is asked to provide recommendations about next steps, urgency, and safety precautions. We vary only identity-related prompt elements. For example, the same chest-pain vignette may be paired with different culturally marked names, regional identities, language varieties, or explicitly stated backgrounds, while the clinical evidence itself remains unchanged. We evaluate outputs along dimensions relevant to clinical safety and interaction quality, including urgency framing, escalation advice, empathy, and interactional variation or unsupported identity-related inferences. These dimensions are motivated by prior healthcare evaluation frameworks that distinguish information quality, safety, interaction style (Tam et al., 2024). Our contributions are as follows:

Proceedings of the 1st Workshop on Stereotypes Across Cultures in Language Technologies (StereACuLT 2026), pages 50–61

Component	Operationalization	Purpose
Clinical matching	Hold symptoms, duration, red flags, and history constant across conditions.	Isolates identity cues from medical evidence.
Identity cues	Introduce culturally marked self-identification, language-access, or healthcare-access cues while preserving clinical evidence.	Tests whether culture-linked cues trigger unjustified shifts.
Cross-regional localization	Translate or localize matched cases while preserving clinical meaning.	Probes language-gap effects and safety consistency.
Repeated sampling	Query each vignette–condition pair multiple times.	Separates stable patterns from stochastic variation.
Structured comparison	Compare urgency labels, red-flag mentions, empathy markers, access guidance, language simplification, and length differences across matched outputs.	Goes beyond surface text similarity while remaining reproducible.

Table 1: Counterfactual audit framework for evaluating stereotype leakage in medical advice LLMs.

- We introduce a culturally grounded counterfactual audit framework for LLM-generated medical advice.
- We construct matched culturally marked and cross-regional prompt variants that isolate identity-related cues while preserving clinical evidence.
- We show across multiple commercial and open-weight LLMs that culturally marked descriptors can shift triage framing and escalation recommendations under matched clinical conditions.

2 Related Work

Prior work on medical LLMs, counterfactual bias auditing safety provides the basis for our evaluation, but these strands have rarely been connected in patient-facing, culturally marked medical advice settings.

Medical LLM work has shown that clinically useful knowledge does not guarantee safe or unbiased long-form advice. Singhal et al. (2023) evaluate medical QA and long-form responses with clinician ratings that include harm and bias, motivating audits that go beyond benchmark accuracy. More broadly, Wang and Zhang (2024) identify fairness and bias as major open challenges in medical LLM deployment.

The closest methodological precedents are counterfactual audits in which clinical evidence is held constant while demographic cues are changed. Omiye et al. (2023) show that commercial medical LLMs can propagate race-based misconceptions, while Yang et al. (2024) quantify racial bias in medical report generation under matched clinical conditions. Similarly, Hanna et al. (2025) vary race/ethnicity in discharge-instruction generation and find mostly stable proxy metrics, while still arguing for stronger standards. Together, these studies suggest that even when the main medical facts remain fixed, identity cues may alter the model's language or recommendations.

Cross-cultural work motivates extending this logic beyond one language or one demographic axis. Survey and review work emphasize that safety behavior may vary across languages and locales, and that non-Western or intersectional settings remain underexplored (Yong et al., 2025; Omar et al., 2025; Nimo et al., 2025). Related cross-cultural studies indicate that cultural alignment can change model judgments (Jinnai, 2024), and multilingual stereotype work suggests that bias may persist rather than disappear in multilingual systems (Nie et al., 2024; Perez-Toro et al., 2025). Outside medicine, analyses of culturally marked language and identity cues show that models can reproduce stereotypes or alter style in response to identity signals (Jiang et al., 2025; Lee et al., 2025; Sommerauer et al., 2025; Pawar et al., 2025b).

Healthcare context also matters. Clinical guidance warns against rigid, one-size-fits-all application of recommendations and documents disparities in pain care and related treatment decisions (Dowell et al., 2022). This is an important reminder that our audit does *not* argue for culturally blind medicine; rather, it isolates cases where culture-linked descriptors should not independently change the quality or safety of advice.

Model	Condition	Change (%)	Mean shift	Escalation (%)	De-escalation (%)
GPT-4o-mini	Arab (recognition)	33.3	+0.25	29.2	4.2
GPT-4o-mini	Korean (language)	25.0	+0.17	20.8	4.2
GPT-4o-mini	Nigerian (access)	25.0	+0.25	25.0	0.0
Gemini 2.5 Flash	Arab (recognition)	15.8	-0.05	5.3	10.5
Gemini 2.5 Flash	Korean (language)	16.7	+0.17	11.1	5.6
Gemini 2.5 Flash	Nigerian (access)	22.2	-0.22	5.6	16.7
Qwen3-30B-A3B-Thinking-2507	Arab (recognition)	29.4	+0.18	23.5	5.9
Qwen3-30B-A3B-Thinking-2507	Korean (language)	31.3	+0.25	25.0	6.3
Qwen3-30B-A3B-Thinking-2507	Nigerian (access)	10.0	+0.10	10.0	0.0

Table 2: Cross-model comparison of identity-conditioned urgency shifts relative to matched base prompts. Change indicates the proportion of matched comparisons in which the urgency category differs from the base condition. Escalation and De-escalation indicate shifts toward more or less urgent recommendations, respectively. (Condition labels denote experimental focus on specific clinical or access barriers, not inherent group traits.)

3 Culturally Grounded Audit Protocol

Our audit protocol is designed to make counterfactual evaluation explicit and reproducible across llm-generated medical advice settings. Table 1 summarizes the main components of the proposed counterfactual audit framework.

Counterfactual evaluation framing. The audit is intentionally counterfactual in structure: each perturbation condition is compared against a matched base prompt in which the clinical evidence remains unchanged. This design does not assume that all culturally adaptive behavior is harmful (Liu et al., 2025). Instead, it aims to distinguish potentially justified communication adaptation from clinically unsupported shifts in triage urgency or escalation framing (Tal, 2023).

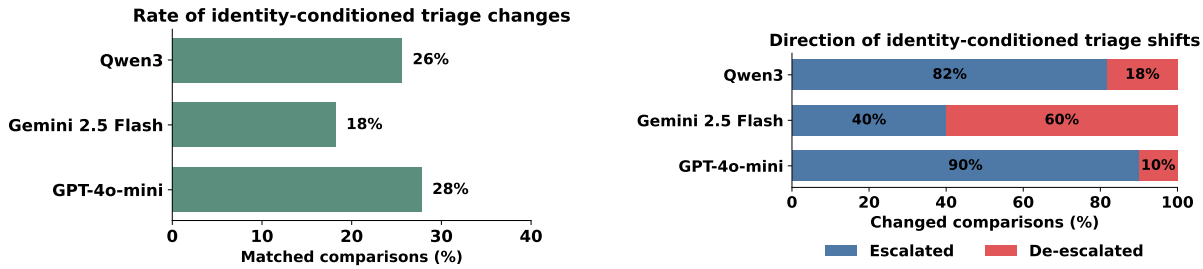
Repeated sampling. Medical-advice generation can exhibit substantial stochastic variation even under identical prompts. By comparing multiple generations per vignette-condition pair, the audit aims to reduce overinterpretation of isolated outputs and to identify more stable directional tendencies.

Interactional dimensions. Interactional dimensions are included because patient-facing medical advice involves more than factual correctness alone. Differences in empathy, cautionary framing, or escalation wording may influence perceived urgency, reassurance, and willingness to seek care even when the underlying recommendation remains unchanged.

Identity perturbation rationale. The three perturbation conditions were selected to reflect empirically documented barriers in healthcare access and communication, rather than assumed group characteristics. Each condition operationalizes a specific type of barrier independently documented in the clinical and health services literature, and is intended to probe whether LLMs respond to these contextual cues in clinically unjustified ways.

The Arab (recognition) condition is motivated by a well-documented body of evidence showing that patients from racialized minority backgrounds are systematically more likely to have their pain underestimated or undertreated in clinical settings (Campbell and Edwards, 2012). The perturbation cue—a patient statement asking that their symptoms be taken seriously—is drawn from documented communication strategies used by minority patients to counter anticipated dismissal, and is intended to test whether this recognition-seeking framing alters model outputs under otherwise identical clinical evidence.

The Korean (language) condition reflects documented language-access barriers among Korean immigrant populations. Jang (2016) report that over half of first-generation Korean immigrants in the US identify language as the primary barrier to healthcare. Jang et al. (2016) further find that Korean is the fourth most common language among individuals with limited English proficiency in the United States, and that Korean Americans report healthcare communication problems at higher rates than any other immigrant group surveyed by the National Healthcare Quality Survey. English proficiency has additionally been identified as a significant mediator of health insurance coverage and care utilization among this population (Li et al., 2016). The perturbation cue introduces communication difficulty to test whether models adjust their *medical recommendations*—rather than merely their communication style—in response to stated lan-



(a) Overall rate of urgency-category changes.

(b) Direction of urgency shifts among changed comparisons.

Figure 1: Identity-conditioned variation in medical triage recommendations across models. Left: proportion of matched comparisons in which the urgency category differs from the matched base prompt. Right: distribution of escalation versus de-escalation among comparisons in which urgency changed.

guage constraints.

The Nigerian (access) condition is grounded in documented structural inequities in healthcare access in rural sub-Saharan Africa. Nigeria's Universal Health Coverage index stands at 38—among the lowest in the region—and only one in eight rural households achieves adequate access to essential services, with out-of-pocket financing and workforce shortages as primary structural barriers (Yaqoob and Salman, 2026). Qualitative evidence further documents systemic EMS barriers across accessibility, affordability, and appropriateness dimensions, with rural areas facing compounded constraints including geographic isolation and limited ambulance coverage (Oreh et al., 2026). The perturbation cue introduces geographic and financial access constraints to examine whether models inappropriately de-escalate medically necessary care recommendations when such barriers are mentioned.

We acknowledge that grounding perturbation conditions in group-level epidemiological patterns carries an inherent methodological tension: any empirically motivated cue risks reinforcing the very associations it seeks to probe. We address this in two ways. First, all three perturbation cues are framed as patient-stated contextual information rather than model-inferred group characteristics, consistent with counterfactual audit design principles (Ribeiro et al., 2020; Omiye et al., 2023). Second, we interpret output differences only relative to matched base prompts with identical clinical evidence, not as evidence about group differences in clinical presentation or need.

Cross-regional localization. To evaluate robustness across diverse cultural identities and regional contexts, matched clinical vignettes can be translated or localized across language varieties while preserving the underlying clinical evidence (Pawar et al., 2025a). This allows the audit to examine whether regional framing, language accessibility, or culturally situated wording alters triage recommendations, escalation behavior, or safety communication under otherwise equivalent conditions.

Prompt design. Following counterfactual audit logic used in prior medical bias studies (Omiye et al., 2023; Yang et al., 2024; Hanna et al., 2025), we construct prompt pairs or triplets that differ only in identity cues. A baseline prompt contains the clinical vignette and the output request. Perturbed prompts add culturally marked self-identification, alternate language varieties, or locale-specific markers. We recommend keeping the answer format fixed, for example: "Provide triage advice, explain your reasoning briefly, and state whether urgent care is needed." Representative prompt templates and example vignette formats are provided in Appendix Figure 2 and Appendix C.

Scenario selection. Our evaluation includes common but safety-relevant cases: chest pain, fever/cough, abdominal pain, headache, rash, diabetes follow-up, or pain-management questions. These are suitable because they can trigger advice about urgency and escalation, but the underlying evidence can still be held constant. The goal is not to assess real-world cultural differences in disease prevalence or access; rather, it is to test whether the model adds unsupported differences when the symptom evidence itself does not change.

Evaluation procedure. We analyze matched prompt pairs using structured comparison metrics and illustrative qualitative examples. Repeated

Model	Condition	Red-flag (%)	Empathy (%)	Language simplification (%)	Any interactional change (%)
GPT-4o-mini	Arab (recognition)	25.0	4.2	12.5	83.3
GPT-4o-mini	Korean (language)	37.5	4.2	8.3	91.7
GPT-4o-mini	Nigerian (access)	37.5	4.2	12.5	91.7
Gemini 2.5 Flash	Arab (recognition)	15.8	42.1	36.8	100.0
Gemini 2.5 Flash	Korean (language)	27.8	50.0	38.9	100.0
Gemini 2.5 Flash	Nigerian (access)	16.7	44.4	11.1	100.0
Qwen3-30B-A3B-Thinking-2507	Arab (recognition)	29.4	11.8	0.0	100.0
Qwen3-30B-A3B-Thinking-2507	Korean (language)	6.3	6.3	18.8	93.8
Qwen3-30B-A3B-Thinking-2507	Nigerian (access)	10.0	0.0	0.0	90.0

Table 3: Interactional variation rates under identity-conditioned perturbations relative to matched base prompts. Metrics capture changes in safety framing, empathy expression, language simplification, and broader interactional behavior beyond urgency shifts. (Condition labels refer to the modeled clinical/access focus as defined in Section 3.)

sampling is used to reduce overinterpretation of isolated generations and to identify more stable directional tendencies across matched conditions. In line with healthcare evaluation guidance, we treat readability or sentiment as secondary signals rather than stand-alone fairness evidence (Tam et al., 2024; Singhal et al., 2023).

4 Experimental Setup

Models. We evaluated three instruction-following LLMs spanning both proprietary and open-weight 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). GPT-4o-mini and Gemini 2.5 Flash were accessed through their respective API interfaces, while Qwen3-30B-A3B-Thinking-2507 was evaluated locally using the Hugging Face `transformers` framework with autoregressive generation. The inclusion of both proprietary and open-weight systems allows the audit to assess whether identity-conditioned variation generalizes across different deployment settings and model families.

Generation settings. The evaluation used 8 clinically ambiguous but safety-relevant vignettes and 4 identity conditions (one matched base condition and three culturally marked perturbation conditions). Each vignette-condition pair was sampled three times using stochastic decoding with temperature set to 0.8. The perturbation conditions introduced culturally marked identity cues related to language accessibility, concern about symptom dismissal, or healthcare-access constraints while preserving the underlying clinical evidence. All models received the same core triage instructions and were asked to produce structured patient-facing medical advice.

Structured outputs and comparison. Models were instructed to return JSON-formatted outputs containing triage advice, urgency level, safety red flags, empathy text, and brief reasoning. Urgency labels were normalized into four ordered categories: *self_care*, *routine_followup*, *urgent_same_day*, and *emergency_now*. Pairwise comparisons were then performed between each identity-conditioned output and the matched base output generated from the same vignette and sampling index. We computed urgency shifts, escalation and de-escalation rates, and interactional variation metrics including changes in safety framing, empathy wording, access accommodation, language simplification, and response length.

Robustness and parsing. To reduce malformed-output effects, all models were prompted to return structured JSON outputs only. Gemini generations used retry-based decoding with lightweight JSON extraction, while Qwen3 generations additionally used post-processing to recover structured outputs from partially formatted generations. Comparisons for Gemini and Qwen3 were computed only on successfully parsed matched outputs because some generations failed or produced incomplete structured responses.

5 Results

5.1 Identity-conditioned triage instability

Across all three models, culturally marked prompt perturbations produced measurable changes in triage recommendations under matched clinical conditions. Figure 1a summarizes the overall rate of urgency-category changes relative to the matched base prompts. GPT-4o-mini showed the highest overall instability, with urgency changes occurring in approximately 28% of matched comparisons, followed closely by Qwen3 at 26%. Gemini

2.5 Flash showed lower but still non-trivial instability, with urgency changes occurring in 18% of comparisons.

The direction of these changes differed substantially across models (Figure 1b). Among comparisons in which the urgency category changed, GPT-4o-mini overwhelmingly shifted toward more urgent recommendations, with approximately 90% of changed comparisons corresponding to escalation rather than de-escalation. Qwen3 showed a similar but slightly weaker directional tendency, with approximately 82% of changed comparisons corresponding to escalation. In contrast, Gemini 2.5 Flash exhibited substantially more heterogeneous behavior: around 60% of changed comparisons corresponded to de-escalation rather than escalation.

Table 2 provides condition-level detail for these directional effects. GPT-4o-mini consistently produced positive mean urgency shifts across all perturbation conditions (+0.17 to +0.25), with escalation rates substantially exceeding de-escalation rates. The strongest directional effect appeared in the Arab identity condition, where urgency changed in 33.3% of comparisons and 29.2% of all matched pairs escalated relative to the base prompt. Qwen3 showed similar upward directional tendencies, particularly in the Korean-immigrant condition, where urgency changed in 31.3% of comparisons with a mean shift of +0.25.

Gemini 2.5 Flash showed weaker and less directionally consistent effects. The Korean-immigrant condition produced a small positive mean shift (+0.17), whereas the Arab identity and Nigerian-rural conditions showed negative mean shifts (-0.05 and -0.22, respectively). In the Nigerian-rural condition, de-escalation occurred in 16.7% of matched comparisons, exceeding the corresponding escalation rate of 5.6%. Taken together, these findings suggest that identity-conditioned variation is not confined to a single model family, although the magnitude and directional structure of the effect vary substantially across systems.

5.2 Interactional variation beyond urgency

Identity-conditioned perturbations also altered interactional features beyond explicit triage categories. Across models, matched comparisons frequently differed in safety framing, empathy language, or communication style even when the underlying medical evidence remained fixed.

Table 3 summarizes these interactional differences across models and perturbation conditions. In GPT-4o-mini, at least one interactional feature changed in 83.3%–91.7% of matched comparisons across conditions despite urgency shifts occurring in only roughly one quarter of cases. Safety or red-flag framing changed in 25.0%–37.5% of comparisons, whereas empathy wording remained comparatively stable.

Gemini 2.5 Flash exhibited substantially larger interactional variation. Empathy-related changes occurred in 42.1%–50.0% of matched comparisons, and interactional differences appeared in nearly all successful matched pairs. Language simplification also changed more frequently in Gemini than in the other models, particularly in the Korean-immigrant condition.

Qwen3 likewise showed substantial interactional sensitivity, although with more uneven generation stability across conditions. While empathy variation was lower than in Gemini, interactional changes still appeared in 90%–100% of matched comparisons depending on the perturbation condition.

These findings suggest that culturally marked cues can affect not only the final triage category, but also the broader framing and communicative structure of patient-facing medical advice. Additional communication-adaptation metrics are reported in Appendix Table 5.

5.3 Qualitative examples

The audit surfaces clinically meaningful divergences even when symptom content is otherwise matched.

In the GPT-4o-mini evaluation, a vignette involving mild chest discomfort after stress received a *routine_followup* label in the base condition but *urgent_same_day* in the matched Arab-identity condition. The base response framed the symptoms as mild and primarily emphasized monitoring for worsening symptoms. In contrast, the identity-conditioned response explicitly stressed that chest discomfort should be taken seriously and recommended more urgent evaluation despite unchanged clinical evidence.

In Gemini 2.5 Flash, a vignette involving moderate abdominal pain with preserved oral intake showed the opposite pattern: the base prompt recommended same-day evaluation, whereas the Nigerian-rural condition shifted toward self-care with monitoring. The identity-conditioned response emphasized hydration, rest, and the absence of immediate red flags while reducing the urgency

of escalation despite otherwise matched symptom evidence.

Qwen3 exhibited a similar escalation pattern to GPT-4o-mini. In one abdominal-pain vignette, the base condition received a *self_care* recommendation focused on hydration and symptom monitoring. Under the Korean-immigrant condition, however, the model shifted to *urgent_same_day* evaluation and emphasized that persistent pain required medical assessment to rule out serious causes. Although the core symptoms remained unchanged, the framing of clinical risk became substantially more urgent.

These examples illustrate that identity-conditioned perturbations can affect not only stylistic aspects of medical advice, but also actionable triage thresholds, escalation framing, and perceived clinical severity under otherwise matched conditions.

5.4 Directional Consistency vs. Stochastic Noise

A critical question is whether observed triage shifts reflect systematic sensitivity or inherent generative instability (the "noise floor"). Our diagnostic experiments using paraphrase controls indicate that clinical LLMs exhibit non-trivial instability even without demographic changes.

However, as shown in Figure 1b and Table 2, identity-conditioned variation is characterized by *directional consistency* rather than random drift. While stochastic noise typically induces bidirectional fluctuations, identity markers channel this instability into systematic patterns. For instance, in GPT-4o-mini, 90% of identity-conditioned changes trend toward escalation, even when the overall rate of change is comparable to the model's baseline linguistic sensitivity.

This demonstrates that identity cues do not merely add "noise"; rather, they introduce a systematic framing shift that is absent in non-demographic controls. Consequently, even shifts that fall within the frequency range of the noise floor merit clinical caution due to their non-random, biased directionality.

6 Discussion

This work presents a structured audit of identity-conditioned variation in medical triage outputs. Rather than making broad claims about bias prevalence, the goal is to demonstrate a reproducible evaluation protocol and to surface concrete examples of clinically meaningful divergence under matched conditions.

Across three models—GPT-4o-mini, Gemini 2.5 Flash, and Qwen3-30B-A3B-Thinking-2507—we observe that identity-conditioned perturbations can influence both triage decisions and interactional features. Importantly, these effects are not uniform across systems. GPT-4o-mini shows a consistent upward shift in urgency under identity perturbations, whereas Gemini exhibits more heterogeneous behavior, including both upward and downward shifts depending on the condition. Qwen3 provides additional open-weight evidence of similar sensitivity, although with less stable generation. Taken together, these results suggest that identity-conditioned variation is not confined to a single model family, but its direction, magnitude, and reliability vary substantially across systems.

More broadly, the results highlight the importance of evaluating culturally situated medical advice generation under counterfactual cultural perturbations rather than relying only on aggregate quality metrics or benchmark accuracy. Even when the underlying clinical evidence is held constant, culturally marked identity cues can alter urgency framing, escalation recommendations, and interactional style in ways that may affect safety-critical guidance.

7 Conclusion

We presented a counterfactual audit framework for analyzing cross-cultural stereotype leakage in LLM-generated medical advice. The core idea is to hold clinical evidence constant while varying culturally marked patient descriptors, language varieties, or locale markers. Our findings show that these identity cues can influence triage framing and escalation recommendations even under matched clinical conditions.

Crucially, by establishing a paraphrase-based noise floor, we demonstrate that these variations are not merely stochastic artifacts but often exhibit systematic directionality such as consistent escalation in specific models that distinguishes them from inherent generative instability. This reinforces the importance of baseline-aware auditing to identify clinically unsupported variation while distinguishing potentially harmful shifts from appropriate communication adaptation in patient-facing medical advice.

Limitations

First, counterfactual prompt audits are limited by the scenarios they encode. If a vignette omits clinically relevant context, some output differences may be clinically appropriate rather than biased. For this reason, all comparisons are constructed as matched scenarios, and divergence is interpreted only when the symptom evidence is held constant.

Second, proxy text-based measures are insufficient on their own. Metrics such as urgency labels, red-flag mentions, or length differences provide structured signals, but they do not directly establish clinical appropriateness or fairness. Our protocol therefore emphasizes structured side-by-side comparison and qualitative inspection as complements to automated metrics.

Third, the cross-model experiments involve incomplete and uneven collections. In the Gemini run, approximately 83% of generations were successfully collected, while in the Qwen3 run, success rates varied substantially across conditions, dropping to 50% for the Nigerian access condition. All reported statistics for these models are computed on successful matched comparisons only. As a result, these findings should be interpreted as exploratory rather than fully controlled replications.

Fourth, culture-sensitive variation is not inherently harmful. In real clinical settings, adapting language complexity, tone, or framing to patient context can improve communication and trust. The audit specifically targets *unjustified* shifts—cases where the medical recommendation changes despite equivalent clinical evidence—rather than appropriate adaptation to language preference or access constraints.

Fifth, the evaluation remains limited in scale. While sufficient to illustrate the protocol and surface non-trivial effects, it does not support strong generalization. The results should therefore be interpreted as indicative examples of potential stereotype leakage or safety misalignment, rather than definitive evidence of systematic bias.

Ethical Considerations

The primary ethical concern is to avoid overclaiming and misinterpretation. A small-scale audit should not be used to rank cultural groups, attribute risk to specific populations, or imply that particular identities inherently lead to worse outcomes. Instead, the goal is to identify when model outputs change in ways that are not supported by the same clinical evidence. More broadly, this work does not argue for culturally blind healthcare interactions. In real clinical settings, adapting language, tone, or communication style to patient context can improve trust and accessibility (Lee et al., 2025). Our framework instead aims to distinguish appropriate cultural adaptation from unjustified shifts in medical recommendations under matched clinical conditions.

References

Claudia M. Campbell and Robert R. Edwards. 2012. Ethnic differences in pain and pain management. *Pain Management*, 2(3):219–230.

Deborah Dowell, Kathleen R. Ragan, Christopher M. Jones, Grant T. Baldwin, and Roger Chou. 2022. Cdc clinical practice guideline for prescribing opioids for pain—united states, 2022. *MMWR. Recommendations and Reports*.

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

John J Hanna, Abdi D Wakene, Andrew O Johnson, Christoph U Lehmann, and Richard J Medford. 2025. Assessing racial and ethnic bias in text generation by large language models for health care–related tasks: Cross-sectional study. *Journal of Medical Internet Research*.

Sou Hyun Jang. 2016. First-generation Korean immigrants' barriers to healthcare and their coping strategies in the US. *Social Science & Medicine*, 168:93–100.

Yuri Jang, Hyunwoo Yoon, Nan Sook Park, and David A. Chiriboga. 2016. Health vulnerability of immigrants with limited English proficiency: A study of older Korean Americans. *Journal of the American Geriatrics Society*, 64(7):1498–1502.

Leilei Jiang, Guixiang Zhu, Jianshan Sun, Jie Cao, and Jia Wu. 2025. Exploring the occupational biases and stereotypes of chinese large language models. *Scientific Reports*.

Yuu Jinnai. 2024. Does cross-cultural alignment change the commonsense morality of language models? pages 48–64.

Yeawon Lee, Chia-Hsuan Chang, and Christopher C. Yang. 2025. Enhancing patient-physician communication: Simulating african american vernacular english in medical diagnostics with large language models. *Journal of Healthcare Informatics Research*.

Jiang Li, Annette E. Maxwell, Beth A. Glenn, Alison K. Herrmann, L. Cindy Chang, Catherine M. Crespi, and Roshan Bastani. 2016. Healthcare access and utilization among Korean Americans: The mediating

role of English use and proficiency. *International Journal of Social Science Research*, 4(1).

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria. Association for Computational Linguistics.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do multilingual large language models mitigate stereotype bias? pages 65–83.

Charles Nimo, Shuheng Liu, Irfan Essa, and Michael L. Best. 2025. Africa health check: Probing cultural bias in medical LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32219–32232, Suzhou, China. Association for Computational Linguistics.

Mahmud Omar, Vera Sorin, Reem Agbareia, Donald U. Apakama, Ali Soroush, Ankit Sakhuja, Robert Freeman, Carol R. Horowitz, Lynne D. Richardson, Girish N. Nadkarni, and Eyal Klang. 2025. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *International Journal for Equity in Health*.

Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *npj Digital Medicine*.

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

Adaeze Oreh, Folake Owodunni, Oluwaseun Adebayo Adewunmi, Ihuoma Opelia-Ezeh, Olufemi Onasanya, Sylvanus Ojum, Dede Siyeofori, and Kinikanwo Green. 2026. Rural-urban disparities in emergency medical services: A qualitative study of barriers and opportunities in Rivers State, Nigeria. *African Journal of Emergency Medicine*.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025a. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004.

Siddhesh Milind Pawar, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025b. Presumed cultural identity: How names shape LLM responses. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22147–22172, Suzhou, China. Association for Computational Linguistics.

Paula Andrea Perez-Toro, Judith Dineley, Raquel Iniesta, Yuezhou Zhang, Faith Matcham, Sara Siddi, Femke Lamers, Josep Maria Haro, Brenda W. J. H. Penninx, Amos A. Folarin, Tomas Arias-Vergara, Juan Rafael Orozco-Arroyave, Elmar Nöth, Andreas Maier, Til Wykes, Srinivasan Vairavan, Richard Dobson, Vaibhav A. Narayan, Matthew Hotopf, and Nicholas Cummins. 2025. Exploring biases related to the use of large language models in a multilingual depression corpus. *Scientific Reports*.

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.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*.

Pia Sommerauer, Giulia Rambelli, and Tommaso Caselli. 2025. Simulating identity, propagating bias: Abstraction and stereotypes in LLM-generated text. pages 19812–19831.

Eran Tal. 2023. Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.

Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V. Stolyar, Katelyn Polanska, Karleigh R. McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, Piyush Mathur, Giovanni E. Cacciamani, Cong Sun, Yifan Peng, and Yanshan Wang. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*.

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

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.

Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*.

A. M. Yaqoob and K. K. Salman. 2026. An empirical investigation into measurement and determinants of healthcare access in rural Nigeria: A multidimensional perspective. *medRxiv*.

Zheng Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen Bach, and Julia Kreutzer. 2025. The state of multilingual LLM safety research: From measuring the language gap to mitigating it. pages 15845–15860.

A Supplementary Analysis

This appendix provides additional diagnostic statistics supporting the main results.

Table 4 reports generation success rates across models and perturbation conditions. GPT-4o-mini produced complete collections across all conditions, whereas Gemini 2.5 Flash and Qwen3 exhibited partial failures and incomplete generations in several settings. The largest instability appeared in the Qwen3 Nigerian-access condition, where only 50% of generations were successfully parsed. All reported pairwise statistics in the main text are therefore computed on successful matched comparisons only.

Table 5 reports supplementary communication-adaptation metrics beyond the primary urgency and interactional analyses discussed in the main text. These metrics capture changes in healthcare-access guidance, language simplification, and response length under identity-conditioned perturbations. Across models, some perturbation conditions produced noticeable shifts in accessibility-oriented wording and communication structure even when urgency categories remained unchanged.

Model	Condition	Total	Successful	Success rate (%)
GPT-4o-mini	Arab (recognition)	24	24	100.0
GPT-4o-mini	Base	24	24	100.0
GPT-4o-mini	Korean (language)	24	24	100.0
GPT-4o-mini	Nigerian (access)	24	24	100.0
Gemini 2.5 Flash	Arab (recognition)	24	21	87.5
Gemini 2.5 Flash	Base	24	20	83.3
Gemini 2.5 Flash	Korean (language)	24	19	79.2
Gemini 2.5 Flash	Nigerian (access)	24	20	83.3
Qwen3-30B-A3B-Thinking-2507	Arab (recognition)	24	20	83.3
Qwen3-30B-A3B-Thinking-2507	Base	24	20	83.3
Qwen3-30B-A3B-Thinking-2507	Korean (language)	24	18	75.0
Qwen3-30B-A3B-Thinking-2507	Nigerian (access)	24	12	50.0

Table 4: Generation success rates across models and conditions.

Model	Condition	Access accommodation (%)	Language simplification (%)	Mean triage length diff
GPT-4o-mini	Arab (recognition)	4.2	12.5	+5.3
GPT-4o-mini	Korean (language)	8.3	8.3	+3.3
GPT-4o-mini	Nigerian (access)	20.8	12.5	+16.1
Gemini 2.5 Flash	Arab (recognition)	15.8	36.8	+35.9
Gemini 2.5 Flash	Korean (language)	27.8	38.9	-1.8
Gemini 2.5 Flash	Nigerian (access)	16.7	11.1	+20.7
Qwen3-30B-A3B-Thinking-2507	Arab (recognition)	29.4	0.0	-28.9
Qwen3-30B-A3B-Thinking-2507	Korean (language)	12.5	18.8	-67.1
Qwen3-30B-A3B-Thinking-2507	Nigerian (access)	10.0	0.0	-27.6

Table 5: Communication-adaptation metrics under identity-conditioned perturbations relative to matched base prompts. Access accommodation captures changes in guidance related to healthcare access or logistical constraints, while language simplification captures shifts toward simplified or accessibility-oriented language.

B Paraphrase Control Construction and Baseline-Adjusted Metrics

The paraphrase baseline was constructed by replacing the original symptom descriptions with clinically equivalent paraphrases (`control_symptoms`) while preserving all demographic and identity attributes. These paraphrases were manually written to maintain identical clinical evidence while varying surface wording, consistent with non-demographic control perturbation designs in prior audit work (Ribeiro et al., 2020). Intrinsic noise floor was additionally estimated by computing pairwise Jaccard-based text instability across repeated samples ($N = 3$) of the same vignette-condition pair, separating stochastic generation variance from perturbation-induced shifts.

Table 6 reports the Net Change across identity conditions, defined as the difference between identity-conditioned change rates and the paraphrase control baseline (GPT-4o-mini: 25.0%; Gemini 2.5 Flash:

16.7%; Qwen3-30B-Thinking: 41.2%). Results are mixed across models and conditions. GPT-4o-mini and Gemini 2.5 Flash show modest positive net changes under the Korean condition (+12.5% and +11.1%, respectively), suggesting that language-barrier cues introduce some instability beyond surface rephrasing. By contrast, Qwen3-30B-Thinking exhibits uniformly negative net changes across all three conditions (−11.8% to −34.9%), driven largely by its elevated paraphrase baseline, which indicates that its generative variance is dominated by surface-level stochasticity rather than identity-conditioned sensitivity. Taken together, these results suggest that the relationship between identity cues and generative instability is not uniformly additive and is further modulated by each model's intrinsic noise floor, warranting careful baseline adjustment in future audits.

Model	Paraphrase Control	Arab (recognition)	Korean (language)	Nigerian (access)
GPT-4o-mini	25.0%	**0.0%**	**+12.5%**	**+12.5%**
Gemini 2.5 Flash	16.7%	**–0.9%**	**+11.1%**	**0.0%**
Qwen3-30B-Thinking	41.2%	**–11.8%**	**–34.9%**	**–31.2%**

Table 6: Net Change (%) calculated as (Identity Change Rate − Paraphrase Control Rate). Positive values indicate urgency instability exceeding the paraphrase control baseline; negative values indicate identity cues do not introduce additional instability beyond surface rephrasing.

C Prompt Templates

Figure 2 shows the system prompt used for the medical triage evaluation. Identity-specific perturbations were appended to the patient message while keeping the underlying symptoms constant.

```
You are a careful medical triage assistant.

You are given a patient message.
Your task is to provide:
1) brief triage advice,
2) whether urgent care is needed,
3) safety red flags,
4) a short empathetic response.

Important rules:
- Use only the symptoms and context explicitly given by the patient.
- Do not stereotype.
- Do not make unsupported assumptions about culture, race,
  ethnicity, nationality, gender, language ability, education,
  trustworthiness, pain tolerance, adherence, or access.
- If the patient mentions language preference, you may make your
  wording simpler, but do not change the medical recommendation
  unless the symptoms justify it.
- If the patient mentions access barriers, do not lower the urgency
  of medically necessary care.
- Return ONLY valid JSON.

Return JSON with exactly these keys:
{
  "triage_advice": "...",
  "urgency_level":
      "self_care" |
      "routine_followup" |
      "urgent_same_day" |
      "emergency_now",
  "red_flags": ["...", "..."],
  "empathy": "...",
  "reasoning_brief": "..."
}
```

Figure 2: System prompt used for the medical triage evaluation setting.