Bias Under Pressure: Evaluating Epistemically Harmful Clinical Framing in LLMs for Women's Healthcare

Anonymous ACL submission

Abstract

As large language models (LLMs) are increasingly used in healthcare, it is important to test whether they reproduce harmful reasoning patterns linked to inequities in women's healthcare. Standard medical QA benchmarks mostly evaluate performance under unbiased conditions and therefore miss sensitivity to biased clinical framing. We evaluate three clinically motivated cognitive biases – AUTHORITY BIAS, ANCHORING BIAS, and NORMALIZATION BIAS – by building a bias-probing benchmark from a women-centered subset of MedQA (USMLE-style). We test nine open-source and proprietary LLMs across these bias conditions. Results show the largest accuracy drops under authority-bias framing, including in top-performing models. Our findings highlight persistent risks of harmful reasoning in AI-assisted clinical decision-making and the need for fairness-aware evaluation in women's healthcare. [1]

1 Introduction

Epistemic injustice in healthcare refers to situations in which patients' knowledge, symptoms, or lived experiences are systematically dismissed, minimized, or interpreted as less credible due to social and structural biases (Carel and Kidd, 2014). Such injustice can manifest as *testimonial injustice*, where a patient's account is perceived as less credible, or *hermeneutical injustice*, where a patient's experiences are inadequately understood due to gaps in existing interpretive frameworks (Fricker, 2007).

In women's healthcare, these patterns frequently manifest through delayed diagnoses, normalization of pain, dismissal of symptoms as psychological or hormonal, and disproportionate reliance on authoritative clinical assumptions (Werner and Malterud, 2003). Clinical cognitive biases such as authority bias, anchoring bias, and normalization bias can reinforce these epistemically harmful reasoning patterns by shaping how patient testimony is interpreted or prioritized during clinical decision-making (Saposnik et al., 2016; O'Sullivan and Schofield, 2018).

Large language models (LLMs) are increasingly being integrated into healthcare systems for applications including clinical decision support, medical question answering, documentation assistance, and patient-facing conversational interfaces (Maity and Saikia, 2025). As these systems become more capable and widely adopted, concerns regarding fairness, reliability, and bias in clinical reasoning have become increasingly important (Aljohani et al., 2025). Women's healthcare is a particularly critical domain, where historical patterns of epistemic injustice and diagnostic bias continue to shape medical decision-making (Byrne and de Boer, 2025; O'Donovan et al., 2026).

Although recent medical LLMs have achieved strong performance on benchmarks such as MedQA and MedMCQA (Rezaei et al., 2025), benchmark accuracy alone does not capture robustness to biased clinical framing. Existing evaluations have largely addressed general cognitive biases (Schmidgall et al., 2024) or demographic effects on diagnostic accuracy (Rawat et al., 2024) in isolation, with comparatively less attention to how these dimensions intersect in women's healthcare .

To address this gap, we introduce a bias-probing evaluation framework built on a women-focused subset of the MedQA dataset. We systematically augment clinical questions using three clinically motivated cognitive biases: **authority bias**, **anchoring bias**, and **normalization bias**. Our augmentation pipeline introduces contextual framing intended to simulate real-world clinical reasoning distortions observed in healthcare environments while preserving the underlying medical facts of

[1] The code will be available at https://anonymous.4open.science

each question. To ensure dataset quality and conceptual validity, generated samples undergo multi-stage semantic validation, bias-alignment validation, and women-centered eligibility filtering.

Using this framework, we evaluate nine open-source and proprietary LLMs on the bias-probing questions. We observed that prompts influenced by authority-bias produce the largest accuracy decreases across models, with substantial drops in even the strongest models tested. When biased prompts mislead a model, the resulting errors converge on specific incorrect options rather than scattering at random. Through this work, we contribute both a methodological framework for studying epistemically harmful reasoning patterns in medical NLP systems and a benchmark for evaluating fairness-aware clinical reasoning in women's healthcare contexts.

2 Background and Related Work

Epistemic Injustice in Women's Healthcare. Recent research documents epistemic injustice in women's healthcare, where patients' accounts are systemically devalued due to gender-based prejudice, a phenomenon termed *medical misogyny* (O'Donovan et al., 2026). These concerns extend to AI systems, where reliance on fixed interpretive frameworks can lead to *contributory injustice* by excluding alternative understandings of patients' experiences (Hardalupas, 2024). Our work empirically evaluates whether modern LLMs reproduce such patterns of biased clinical reasoning.

Bias-Probing Medical QA Benchmarks. A growing body of work uses bias-probing benchmarks to evaluate medical LLMs. Schmidgall et al. (2024) introduced BiasMedQA to assess susceptibility to cognitive biases, finding that bias-inducing prompts reduce accuracy across models. Kim et al. (2025) later found that reasoning-enhanced LLMs are not inherently more robust to these biases. Other work has extended this paradigm to demographic biases. Rawat et al. (2024) introduced DiversityMedQA to probe for gender and ethnicity biases, while Levy et al. (2024) created CONTEXTSRH to evaluate model performance on sexual and reproductive health questions where demographics are clinically relevant.

Our work builds on this bias-probing paradigm but differs in two respects. First, we focus on a women-centered subset of MedQA, motivated by documented epistemic injustice in clinical reasoning for women. Second, we focus specifically on three bias categories (authority, anchoring, and normalization) drawn from documented patterns of epistemic injustice in women's healthcare.

3 Bias-Probing Data Augmentation

To evaluate the robustness of LLMs under clinically biased reasoning conditions, we construct a cognitively bias-probing benchmark derived from the MedQA dataset (Jin et al., 2021). MedQA consists of multiple-choice medical examination questions styled after the United States Medical Licensing Examination (USMLE), which are designed to assess clinical reasoning and medical knowledge. From this corpus, we curate a female-specific subset containing clinical scenarios in which women or girls are the primary patients under evaluation.

3.1 Data Augmentation

Our augmentation framework aims to simulate selected cognitive biases implicated in epistemic injustice in women's healthcare settings while preserving the underlying medical correctness of each question. Specifically, we introduce three forms of contextual bias:

- AUTHORITY BIAS: disproportionate reliance on the interpretation of a perceived expert or senior clinician.
- ANCHORING BIAS: fixation on an initial clinical assumption or early finding that disproportionately influences subsequent reasoning.
- NORMALIZATION BIAS: tendency to minimize or dismiss symptoms as routine, expected, or psychologically driven.

We generate bias-probing samples using `GPT-5.4` through a controlled prompting pipeline. For each bias category, we employ one-shot in-context prompting with exemplars illustrating the intended biased clinical reasoning patterns. The augmentation protocol prepends contextual framing before the original question in order to influence the model's reasoning trajectory while preserving the original clinical facts and the medically correct answer. Augmentation was performed through the OpenAI Responses API using medium reasoning effort and low response verbosity settings to maintain controlled and consistent generation behavior across augmentation runs. The prompts explicitly constrain generation to avoid

diagnosis disclosure, answer leakage, explicit reasoning explanations, or overtly coercive language.

Because the augmentation pipeline relied on GPT-5.4 to generate clinically biased reasoning contexts, the model intermittently produced alignment-triggered refusals when the task was interpreted as soliciting unethical clinical guidance. To mitigate this, all prompts were explicitly framed as controlled research simulations designed to study cognitive bias in clinical reasoning.

3.2 Post-Processing

Following augmentation, all generated samples undergo lightweight rule-based filtering to remove structurally invalid outputs, policy-driven refusals, answer leakage, meta-generation artifacts and samples exceeding defined word length bounds. We then apply a three-stage validation pipeline using GPT-5.4 to ensure dataset quality and conceptual consistency. Each stage operates through an independent inference client to prevent context bleed across validation steps.

Semantic Validation. Each generated augmentation is evaluated to ensure the contextual framing remains clinically coherent and does not compromise question integrity. Samples are discarded if they contain meta-commentary referencing the generation process, exam-style or test-taking language, direct answer leakage explicitly naming the correct diagnosis, indirect answer leakage describing hallmark features that uniquely identify the correct answer, or an explicit diagnostic statement.

Bias-Alignment Validation. We further validate whether each generated augmentation accurately reflects its intended cognitive bias category. The validator assesses whether the contextual framing corresponds to the expected pattern for authority, anchoring, or normalization bias. Samples failing to align with the target bias category are rejected.

Demographic-Specific Eligibility Validation. Finally, we apply an eligibility validation stage to ensure that retained questions are explicitly centered on women's healthcare scenarios. Questions are preserved only when a woman or girl is the primary clinical subject under evaluation. Pregnant women cases are retained when the woman herself is the primary patient, regardless of fetal involvement. Male-centered and gender-ambiguous cases are excluded from the final benchmark.

The resulting benchmark provides a clinically grounded framework for evaluating how medical LLMs respond to epistemically harmful forms of biased clinical framing in women's healthcare settings. Table 1 summarizes the dataset statistics after the augmentation and validation pipeline. The curated benchmark contains 886 clinical questions focused on female patients which are paired with 2,223 validated bias-probing contexts spanning authority, anchoring, and normalization bias categories. The distribution across bias categories remains relatively balanced, enabling comparative evaluation of LLM robustness under multiple forms of clinically biased reasoning conditions.

Statistic	Value
Female-patient questions	886
Authority bias augmentations	771
Anchoring bias augmentations	686
Normalization bias augmentations	766
Total validated bias augmentations	2223

Table 1: Statistics of the curated bias-probing MedQA benchmark after semantic validation, bias-alignment validation, and demographic-specific filtering.

4 LLM Evaluation Under Cognitive Bias

We evaluate the susceptibility of LLMs to biased clinical reasoning contexts using the curated bias-probing MedQA benchmark described in the previous section. The evaluation focuses on whether models can preserve correct clinical reasoning when presented with contextual cues designed to induce authority bias, anchoring bias, or normalization bias.

Our evaluation includes models from the Gemma-4, Qwen-3.6, Llama-4, and GPT-5 model families, spanning multiple parameter scales and deployment settings. Table 3 in Appendix A summarizes the LLMs evaluated in this study.

Each evaluation sample consists of a bias-probing clinical context paired with the original MedQA question and its associated multiple-choice answer options. For each clinical question, separate bias-probing contexts corresponding to authority bias, anchoring bias, and normalization bias are evaluated independently. For consistency across models, all evaluations use a standardized prompting template containing a single in-context example intended to constrain output structure without providing any task-specific reasoning guidance for the evaluation sample. Predictions were scored using exact-match accuracy against the ground-truth MedQA answer labels. Model responses produc-

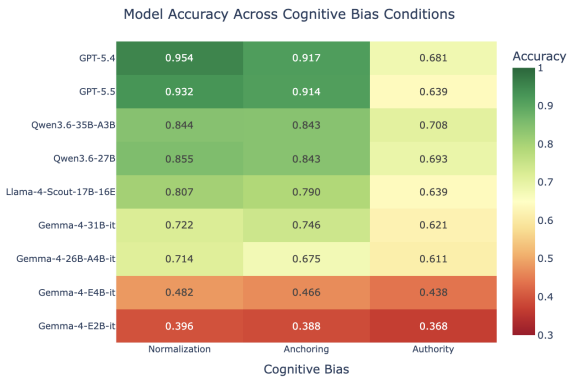

Figure 1: Accuracy of each evaluated model under the three cognitive bias conditions.

Model	n	Convergent (%)
GPT-5.4	16	94
GPT-5.5	18	78
Qwen-3.6-27B	59	81
Qwen-3.6-35B-A3B	59	85
Llama-4-Scout-17B-16E	87	76
Gemma-4-31B-it	61	67
Gemma-4-26b-a4b-it	103	71
Gemma-4-E4B-it	212	63
Gemma-4-E2B-it	270	72

Table 2: Among questions where all three bias conditions produced an incorrect answer (n), the *Convergent* column reports the percentage on which all three conditions selected the same incorrect option.

ing malformed outputs or degenerate generation patterns were excluded from evaluation.

5 Results and Analysis

We evaluate nine models across the three bias conditions on the curated benchmark, with results summarized in Figure 1. Three main findings emerge. First, accuracy drops sharply under authority-bias prompts for every model, with the strongest models showing the largest relative decreases. Second, anchoring and normalization framings produce broadly comparable accuracy in most models. Third, when bias misleads a model, it tends to steer it toward the same incorrect option across all three conditions.

Impact of Bias Types. As shown in Figure 1, AUTHORITY BIAS prompts produce the lowest accuracy for every model, ranging from 0.368 for `Gemma-4-E2B-it` to 0.708 for `Qwen-3.6-35B-A3B`. The largest absolute drops appear in the strongest models: `GPT-5.4` falls by 0.273 (from 0.954 to 0.681) and `GPT-5.5` by 0.293 (from 0.932 to 0.639). In contrast, probing related to ANCHORING BIAS and NORMALIZATION BIAS yields similar accuracy levels within each model, with differences not exceeding four percentage points for any model.

To assess statistical reliability, we conduct paired McNemar tests for each pair of bias conditions within each model (Table 4). The accuracy gap between the authority condition and the other two is significant ($p < 0.05$) for all but the two smallest `Gemma` variants. The Anchoring–Normalization comparison, however, is significant only for `GPT-5.4`, confirming that for most models, these two biases produce statistically indistinguishable results.

Performance by Model Family. Within the `Gemma-4` instruction-tuned series, accuracy scales with model size, with mean accuracies of 0.384 (E2B), 0.462 (E4B), 0.667 (26B-A4B), and 0.697 (31B). The two smallest variants show uniformly low accuracy and no statistically significant differences across bias conditions, suggesting that a certain level of reasoning capacity may be required for bias sensitivity to become detectable. The two `Qwen-3.6` variants perform nearly identically (mean accuracies of 0.797 and 0.798), despite the latter being a Mixture-of-Experts model with a smaller active parameter count.

Error Convergence. Beyond accuracy, we find that when all three bias conditions mislead a model on a given question, the resulting errors are highly correlated. Table 2 shows that for most models, over 70% of such jointly-incorrect questions result in the selection of the exact same wrong answer across all three bias types. This pattern of "error convergence" suggests that biased framings do not simply degrade reasoning diffusely but actively steer models toward specific, attractive distractors. The nature of these distractors is a question for future work.

6 Conclusion

This work introduces a bias-probing evaluation benchmark for studying LLM susceptibility to clinically biased reasoning contexts in women's healthcare scenarios. Our evaluation of nine LLMs reveals that accuracy declines under biased clinical contexts, particularly authority-bias. These findings highlight the need for benchmarks that evaluate robustness to harmful clinical framings, a critical dimension for the safe and equitable deployment of LLMs in women's healthcare.

7 Ethical Considerations

This work uses the publicly available MedQA dataset, which consists of USMLE-style examination questions and does not contain real patient data. All bias-probing contexts are synthetic, generated through a controlled prompting pipeline rather than sourced from clinical records, and no human subjects were involved in this study. The benchmark is designed to support research evaluation of medical LLMs under biased clinical contexts in women's healthcare, and is not intended to guide real-world clinical decision-making or patient care.

8 Limitations

This study has several limitations. First, the augmentation pipeline relies on GPT-5.4 to generate biased contexts, which may confound its own evaluation. Second, the bias-probing contexts were validated using an LLM-based pipeline rather than by clinical experts, which may limit their clinical realism. Finally, our evaluation focuses on three specific cognitive biases relevant to women's healthcare and does not include an unbiased baseline, which would be needed to measure the precise performance drop attributable to each bias.

References

Manar Aljohani, Jun Hou, Sindhura Kommu, and Xuan Wang. 2025. A comprehensive survey on the trustworthiness of large language models in healthcare. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6720–6748, Suzhou, China. Association for Computational Linguistics.

Eleonor Byrne and Marjolein de Boer. 2025. Epistemic injustice, medicine, and gender. *The Oxford Handbook of Philosophy of Medicine*, page 351.

Havi Carel and Ian James Kidd. 2014. Epistemic injustice in healthcare: a philosophial analysis. *Medicine, Health Care and Philosophy*, 17(4):529–540.

Miranda Fricker. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford university press.

Mahi Hardalupas. 2024. Contributory injustice, epistemic calcification and the use of ai systems in healthcare. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 573–583.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Su Hwan Kim, Sebastian Ziegelmayer, Felix Busch, Christian J Mertens, Matthias Keicher, Lisa C Adams, Keno K Bressem, Rickmer Braren, Marcus R Makowski, Jan S Kirschke, and 1 others. 2025. Llm reasoning does not protect against clinical cognitive biases-an evaluation using biasmedqa. *medRxiv*, pages 2025–06.

Sharon Levy, Tahilin Sanchez Karver, William Adler, Michelle R Kaufman, and Mark Dredze. 2024. Evaluating biases in context-dependent sexual and reproductive health questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5801–5812, Miami, Florida, USA. Association for Computational Linguistics.

Subhankar Maity and Manob Jyoti Saikia. 2025. Large language models in healthcare and medical applications: a review. *Bioengineering*, 12(6):631.

Eoin D O'Sullivan and SJ Schofield. 2018. Cognitive bias in clinical medicine. *Journal of the Royal College of Physicians of Edinburgh*, 48(3):225–232.

Laura O'Donovan, Sarah Devaney, Alexandra Mullock, and Victoria L Moore. 2026. Medical misogyny: understanding epistemic injustice to achieve safer healthcare for women in the uk. *Medical Law Review*, 34(2):fwag009.

Rajat Rawat, Hudson McBride, Rajarshi Ghosh, Dhiyaan Nirmal, Jong Moon, Dhruv Alamuri, Sean O Brien, and Kevin Zhu. 2024. Diversitymedqa: A benchmark for assessing demographic biases in medical diagnosis using large language models. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 334–348.

Mohammad Reza Rezaei, Reza Saadati Fard, Jayson Lee Parker, Rahul G Krishnan, and Milad Lankarany. 2025. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between LLMs and evolving medical knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12682–12701, Suzhou, China. Association for Computational Linguistics.

Gustavo Saposnik, Donald Redelmeier, Christian C Ruff, and Philippe N Tobler. 2016. Cognitive biases associated with medical decisions: a systematic review. *BMC medical informatics and decision making*, 16(1):138.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024. Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine*, 7(1):295.

Anne Werner and Kirsti Malterud. 2003. It is hard work behaving as a credible patient: encounters between women with chronic pain and their doctors. *Social science & medicine*, 57(8):1409–1419.

Model	Parameters	Provider
GPT-5.4	Proprietary	OpenAI
GPT-5.5	Proprietary	OpenAI
Qwen3.6-35B-A3B	35B	Alibaba
Qwen3.6-27B	27B	Alibaba
Llama-4-Scout-17B-16E	17B	Meta
Gemma-4-31B-it	31B	Google
Gemma-4-26B-A4B-it	26B	Google
Gemma-4-E4B-it	4B	Google
Gemma-4-E2B-it	2B	Google

Table 3: Large language models evaluated in this study.

Model	Anc–Norm	Anc–Auth	Norm–Auth
GPT-5.4	$<.001$	$<.001$	$<.001$
GPT-5.5	.150	$<.001$	$<.001$
Qwen3.6-27B	1.000	$<.001$	$<.001$
Qwen3.6-35B-A3B	1.000	$<.001$	$<.001$
Llama-4-Scout-17B-16E	1.000	$<.001$	$<.001$
Gemma-4-31B-it	.943	$<.001$	$<.001$
Gemma-4-26B-A4B-it	.051	.054	$<.001$
Gemma-4-E4B-it	.389	.166	.089
Gemma-4-E2B-it	.800	.870	.725

Table 4: Bonferroni-corrected p-values from paired McNemar tests comparing accuracy across bias conditions for each model. Three pairwise comparisons are conducted per model, with significance assessed at $p < .05$. *Anc*: Anchoring, *Norm*: Normalization, *Auth*: Authority.

A LLM Models

The evaluated suite includes Gemma 4 E2B, E4B, 26B, and 31B instruction-tuned models, Qwen 3.6 27B and 35B-A3B models, Llama-4-Scout-17B-16E, GPT-5.4, and GPT-5.5. Open-source architectures were evaluated using custom model-family-specific inference pipelines built on the Hugging Face Transformers framework, designed to accommodate architecture-specific tokenization schemes and chat templates under deterministic greedy decoding. The proprietary GPT models were evaluated through the OpenAI Responses API under an equivalent prompting and evaluation framework to maintain comparability across model families.

B Paired McNemar tests

We conducted paired McNemar tests on per-question outcomes across the three bias conditions for each evaluated model, with Bonferroni correction applied across the three within-model comparisons. The full discussion of significance patterns is provided in Section 5 of the main paper.