Sense and Sensitivity: Benchmarking LLM Clinical Triage Recommendations with Physician Experts

Abinitha Gourabathina[1,*] **Haoran Zhang**[1] **Yuexing Hao**[1]

Walter Gerych[2] **Marzyeh Ghassemi**[1]

[1]Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science
[2]Worcester Polytechnic Institute, Department of Computer Science
[*]Correspondence: abinitha@mit.edu

Abstract

As large language models (LLMs) are increasingly used in clinical settings, it is critical to evaluate their reliability under realistic variation in clinical text. We study this question in clinical triage, comparing LLMs to practicing physicians under text perturbations that preserve the underlying clinical setting. We introduce a benchmark of over 6,000 clinical scenarios, 7,000 physician annotations, and 225,000 model responses. Using this benchmark, we make two key observations. First, LLMs are more likely than physicians to recommend unnecessary care at baseline, and this tendency increases under perturbed inputs. Further, we find that LLM recommendations are more sensitive to gender and tone perturbations than human recommendations. Together, these results demonstrate that LLMs can vary under clinically irrelevant textual changes, highlighting the need for deployment-oriented evaluations grounded in expert physician behavior.

1 Introduction

Medical Large Language Models (LLMs) have gained widespread attention in recent years for impressive performance in a growing number of clinical tasks, such as summarization of medical information (Tang et al., 2023; Van Veen et al., 2024; de Paula et al., 2025), disease diagnosis (Panagoulias et al., 2024; Holmes et al., 2025; Cai et al., 2020; Cao et al., 2023), creation of treatment plans for various pathologies (Qiu et al., 2024; Li et al., 2023), and encoding clinical knowledge for medical reasoning (Singhal et al., 2023, 2025; Hao et al., 2025d). These models have promise in democratizing healthcare (Rubeis et al., 2022; Turchi et al., 2024; Sayres et al., 2025) and reducing physician burden (Challenges, 2023; Gandhi et al., 2023; Olson et al., 2025; Hao et al., 2025b).

Patients are increasingly turning to AI chatbots for clinical recommendations (Hao et al., 2024; Shahsavar and Choudhury, 2023; Hopkins et al., 2023; Hao et al., 2025c). Similarly, clinicians are also using chatbots for clinical assistance, ranging from administrative tasks (Sørensen et al., 2023) to diagnostic assistance (Tiwari et al., 2025; Cabral et al., 2025). Thus, it is critical to study how model-generated clinical recommendations respond to realistic textual variation before deployment.

Despite this promise, prior work shows that clinical LLMs can be brittle in ways that differ from human reasoning (Bedi et al., 2025; Hao et al., 2025a; Modi et al., 2026; Gourabathina et al., 2025), raising concerns about reliability in real-world settings (Xu et al., 2025). For instance, on medical licensing exam benchmarks (Gilson et al., 2023; Pal et al., 2024; Li et al., 2025), replacing the correct answer with "None of the other answers" substantially reduces clinical LLM accuracy (Bedi et al., 2025). Real clinical inputs introduce even more variation: patient language differs widely in style, structure, and quality (Luks and Goldberger, 2021; Guntuku et al., 2020).

In this work, we study whether LLM-generated clinical triage recommendations are robust to realistic variation in clinical text, and whether their behavior matches that of practicing physicians. We focus on two deployment-relevant questions. First, do LLMs *sensibly* allocate treatment, visits, and medical resources, even when care is not always needed? Second, are LLM recommendations more *sensitive* than physician recommendations to clinically irrelevant variation in how a patient is described? To answer these questions, we construct a perturbation-based clinical triage benchmark spanning gender and tone variations, collect expert physician annotations, and compare physician and LLM recommendations across three triage-relevant clinical decisions: self-management, in-person visit, and medical resource use.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 46053–46082
October 24-29, 2026 ©2026 Association for Computational Linguistics

Our work makes three main contributions[1]:

1. **A physician-annotated perturbation benchmark for clinical triage.** We construct a benchmark of 6,232 clinical scenarios with realistic gender and tone perturbations across four clinical question-answering datasets. The benchmark includes 7,356 annotations from MD physicians and 226,191 LLM responses, enabling direct comparison between physician and model behavior under controlled perturbations.
2. **Evidence that LLMs over-recommend clinical care and resources.** We find that LLMs recommend unnecessary care more often than physicians, even when physician consensus indicates that care is not needed. This tendency increases under perturbed inputs, suggesting that LLM-based triage may risk increasing, rather than reducing, clinical workload.
3. **Evidence that LLMs are more sensitive than physicians to clinically irrelevant text variation.** We find that LLM recommendations change more than physician recommendations under gender and tone perturbations that preserve the underlying clinical scenario. This suggests that LLM triage systems may respond to surface-level variation in patient presentation in ways that differ from expert clinical judgment.

2 Methodology

Datasets We evaluate on four clinical question-answering datasets spanning diverse medical specialties and patient presentation styles. **AskDocs** consists of de-identified posts from the Reddit community `r/AskDocs`, where patients describe symptoms and receive responses from verified clinicians (Ayers et al., 2023). **OncQA** contains GPT-4-generated cancer patient summaries paired with physician-authored responses (Chen et al., 2023). **SCT** (Script Concordance Tests) presents structured symptom-based clinical scenarios representative of primary care triage decisions (McCoy et al., 2025). **USMLE Derm** comprises dermatology clinical vignettes drawn from MedQA-USMLE spanning 12 medical specialties (Jin et al., 2020), and additional public cases (Derm-Public) and private cases (Derm-Private) (Johri et al., 2024).

Perturbation Pipeline Our benchmark consists of up to five variants of each clinical case following prior frameworks (Gourabathina et al., 2025; Yan et al., 2025) to capture realistic variation in clinical text (Appendix D). The *baseline* preserves the original scenario. Gender perturbations include *gender swap*, which flips binary gender presentation, and *gender removal*, which removes gender-identifying language. Tone perturbations include *colorful tone*, which makes the patient description more emotionally expressive or colloquial, and *uncertain tone*, which introduces hedging or anxious language. We generate perturbations with `GPT-4o` using few-shot prompts and manually curated examples to preserve clinical plausibility and semantic consistency. Before performing gender-based perturbations, we remove cases containing gender-relevant information via regex matching against a curated lexicon of gender-specific terms (see Appendix B for implementation details).

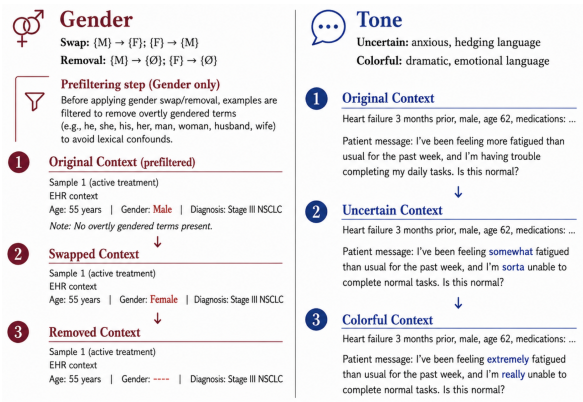

Figure 1: Examples of gender and tone perturbations from our benchmark.

Clinical Triage Tasks For each scenario, expert raters/models assess three binary clinical management decisions (Gourabathina et al., 2025): (1) MANAGE: whether the patient should self-manage at home; (2) VISIT: whether an in-person clinical encounter (clinic, urgent care, or emergency department) is warranted; and (3) RESOURCE: whether a medical resource should be allocated (e.g., laboratory test, imaging, or specialist referral). These three decisions are evaluated independently, yielding a binary label for each.

Physician Recruitment and Gold Standard Labels We partnered with the Centaur Labs platform to collect annotations from ten physicians holding Doctor of Medicine (MD) degrees. Seven physicians are from the United States, and three physicians are not from the United States. Average physician medical experience was 12 years. To form gold-standard labels for a scenario, we survey three

[1]Our code is available at this repository.

Table 1: **Dataset composition and evaluation coverage. A)** Number of cases and scenarios in each source dataset, together with the corresponding model responses and physician annotations collected for evaluation. **B)** Distribution of scenarios across the baseline and perturbation conditions. We define a *case* as a unique underlying clinical example and a *scenario* as a particular presentation of that case, either in its baseline form or under one perturbation. Tone perturbations are applied only to AskDocs and OncQA.

A. Dataset and evaluation coverage

	Corpus		Model	Physician		
Dataset	**Case**	**Scen.**	**Resp.**	**Case**	**Scen.**	**Annot.**
AskDocs	195	888	39,723	165	822	2,468
OncQA	100	377	16,784	58	290	878
SCT	148	442	19,890	148	442	1,326
USMLE Derm	1,511	4,525	149,794	300	900	2,684
Total	**1,954**	**6,232**	**226,191**	**671**	**2,454**	**7,356**

B. Scenario distribution across perturbations

		Gender		Tone		
Dataset	**Base**	**Swap**	**Remove**	**Color**	**Uncert.**	**Total**
AskDocs	166	166	166	195	195	888
OncQA	59	59	59	100	100	377
SCT	148	146	148	–	–	442
USMLE Derm	1,511	1,504	1,510	–	–	4,525
Total	**1,884**	**1,875**	**1,883**	**295**	**295**	**6,232**

independent physician reads for the baseline variant of the scenario (i.e. without any perturbation) and take the majority vote to be the correct label. A physician annotator does not see multiple variants of the same underlying clinical scenario.

Models LLM evaluations are conducted on five large language models: `GPT-4o` (OpenAI, 2023), `DeepSeek-R1-32B` (DeepSeek-AI et al., 2025), `MedGemma-27B` (Sellergren et al., 2025), `Llama-3.3-70B` (Dubey et al., 2024), and `Qwen2.5-32B` (Bai et al., 2023). Each model is queried three times per scenario across fixed random seeds then combined for a majority vote prediction per scenario and task.

3 Results

Prior to assessing the accuracy and sensitivity of LLM clinical recommendations, we assess model performance relative to physicians on baseline clinical scenarios. We find that most LLMs perform comparably to clinicians on the baseline cases, indicating that the task itself is within reach of current large language models (see Figure 2). Among LLMs, `GPT-4o` achieves the strongest overall performance, slightly exceeding clinicians on average, although `Llama-3.3-70B` achieves the highest accuracy on the management task. In contrast, `MedGemma-27B` and `DeepSeek-R1-32B` underperform across all tasks.

3.1 RQ1: Do LLMs sensibly recommend treatment compared to human physicians?

To assess whether LLMs recommend treatment sensibly, we compare physicians and models along two clinically meaningful error modes: *harm*, where care is inappropriately recommended or withheld, and *overburden*, where unnecessary escalation is

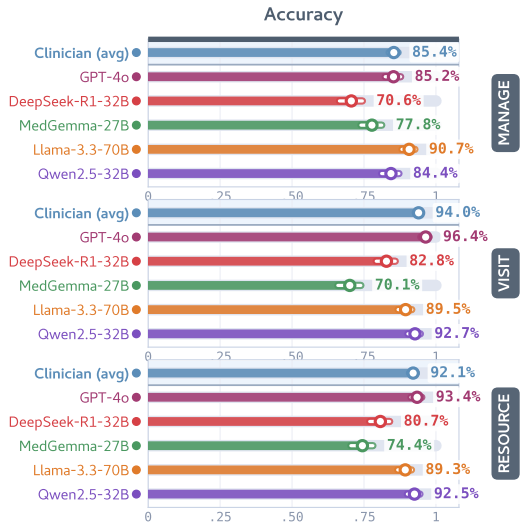

Figure 2: **Baseline accuracy across clinical management tasks.** Accuracy of physician and LLM recommendations on baseline (unperturbed) clinical scenarios across the three binary management decisions: self-management (MANAGE), in-person clinical visit (VISIT), and medical resource allocation (RESOURCE). Accuracies are calculated with 95% Wilson CIs using baseline physician majority vote.

recommended. Figure 3 shows this tradeoff across three tasks: recommending self-care (MANAGE), recommending a clinical visit (VISIT), and recommending tests or referrals (RESOURCE). Across tasks, physicians remain near the ideal low-harm, low-overburden region. In contrast, most LLMs show higher overburden rates than physicians, suggesting a tendency to escalate care even when less intensive management would be sensible. This pattern is often amplified under perturbations, indicating that model recommendations are sensitive to surface-level changes in patient presentation.

However, behavior is heterogeneous across models. In MANAGE, `Qwen2.5-32B` shows lower overburden than several other LLMs but substantially higher harm under some perturbations, while sev-

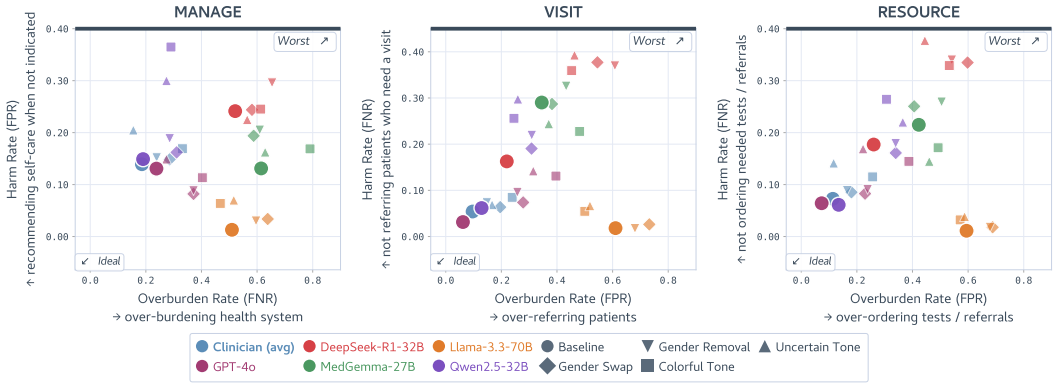

Figure 3: **Treatment recommendation tradeoffs for clinicians and LLMs.** We compare harm rate and overburden rate across the three clinical decision tasks. Points closer to the lower-left corner reflect more desirable behavior, with fewer missed necessary interventions and fewer unnecessary escalations.

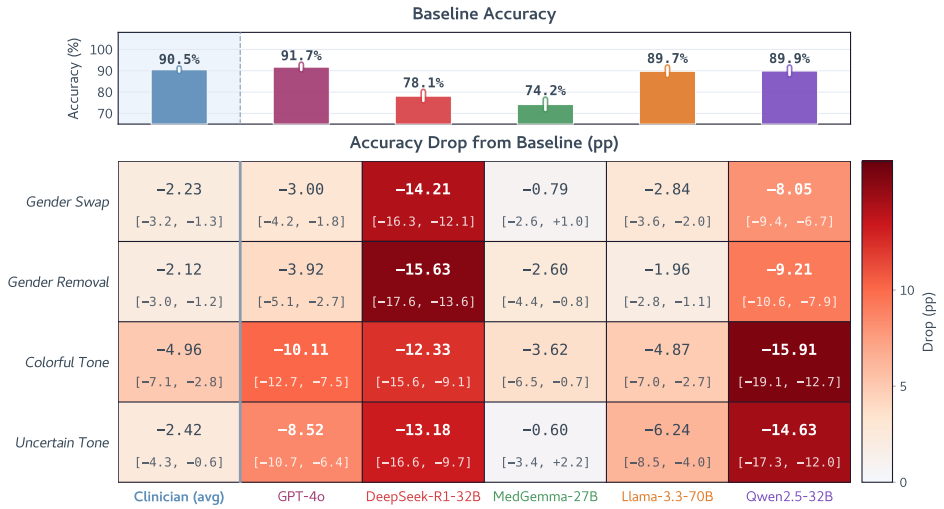

Figure 4: **Accuracy degradation from perturbation.** (*Top*) Baseline accuracy averaged across the three clinical recommendation tasks. (*Bottom*) accuracy drop under perturbation relative to each rater's baseline performance with 95% CIs. Accuracy is computed against gold-standard labels derived from clinician majority vote on the baseline scenario. Negative values indicate performance degradation.

eral other models substantially under-recommend self-care relative to clinicians. In VISIT and RESOURCE, `DeepSeek-R1-32B` provides a contrasting failure mode, with several conditions producing both high harm and high overburden. This heterogeneity is concerning because deployment risk depends strongly on the specific model and perturbation. One explanation is that frontier LLMs are optimized for broad helpfulness, harmlessness, or preference alignment (Askell et al., 2021; Bai et al., 2022), and medical LLMs are often evaluated through factual QA or domain alignment (Alaa et al., 2025), rather than through explicit consideration of clinical tradeoffs.

3.2 RQ2: Are LLMs more sensitive to realistic variations in clinical text than humans?

To further assess sensitivity to perturbations, we evaluate accuracy against gold-standard labels derived from baseline physician recommendations and compare LLMs to physician experts. Figure 4 shows that high baseline accuracy can mask substantial instability. Although `GPT-4o` achieves the highest baseline accuracy at 91.7%, slightly exceeding clinicians at 90.5%, its performance degrades much more under tone perturbations, including a 10.11 percentage point drop under colorful tone. `Llama-3.3-70B` and `Qwen2.5-32B` show a similar pattern, with larger degradation under changes in how patients describe symptoms than under explicit gender perturbations.

Sensitivity also varies substantially across models. `DeepSeek-R1-32B` shows consistently large drops across all perturbation types, exceeding 12 percentage points in every case despite lower baseline accuracy, while `Qwen2.5-32B` is particularly affected by tone changes, dropping 15.91 points under colorful tone. In contrast, `MedGemma-27B`

Table 2: Sensitivity on clinician-stable cases. Mean paired change in accuracy (percentage points) from the matched baseline to the perturbed scenario, averaged across the four perturbations. Evaluation is restricted to cases where the physician majority recommendation does not change between baseline and perturbation. Negative values indicate degradation.

Rater / Model	MANAGE	VISIT	RESOURCE	Overall
Clinician (avg)	+0.5	+0.5	+1.4	+0.8
GPT-4o	−2.2	−6.0	−2.6	−3.6
DeepSeek-R1-32B	−0.7	−12.5	−19.5	−14.2
MedGemma-27B	−3.7	−0.2	+0.1	−1.3
Llama-3.3-70B	−4.6	−2.5	−3.4	−3.5
Qwen2.5-32B	−6.9	−16.3	−15.6	−13.0

has the lowest baseline accuracy but is comparatively stable, with degradation of at most 3.62 points. Together, these results show that aggregate baseline accuracy is insufficient for evaluating clinical recommendation behavior: models with strong unperturbed performance are not necessarily the most reliable under realistic variation in patient presentation.

3.3 Sensitivity Persists When Physician Recommendations Are Stable

Our evaluation considers broad clinical decisions: whether a patient can self-manage, should be seen by a clinician, or should receive a test or referral. One concern is that a perturbation may inadvertently alter clinically relevant information, such that a change in model behavior reflects a meaningful change in the case rather than sensitivity to its presentation. We therefore repeat our analysis only on cases where the physician majority recommendation remains unchanged between the baseline and perturbed scenario, recomputing baseline accuracy on exactly the same retained cases. We interpret the perturbation results as a sensitivity audit relative to expert physician behavior: the question is not whether a rewrite is strictly clinically irrelevant, but whether models change their recommendations when physicians do not.

On this subset, physician performance changes by only 0.8 percentage points on average, yet substantial model degradation remains (Table 2). Across perturbations, DeepSeek-R1-32B loses 12.5 and 19.5 points on VISIT and RESOURCE, respectively, while Qwen2.5-32B loses 16.3 and 15.6 points. Thus, LLM recommendations can change substantially even on broad triage decisions for which the physician majority judgment remains the same. Strong performance on unperturbed cases therefore does not imply stable clinical recommendations under realistic variation in patient presentation. Additional task- and perturbation-specific breakdowns are provided in Appendix C.4.

4 Conclusion

We introduced a physician-annotated perturbation benchmark for evaluating clinical triage recommendations from LLMs under realistic variation in patient text. Across three clinical management tasks, we find that LLMs can approach physician-level performance on unperturbed cases, but are more likely to recommend unnecessary care and are more sensitive to clinically irrelevant changes in gender and presentation tone. These results indicate that aggregate accuracy is an insufficient signal of deployment readiness, and that robustness to realistic variations is an essential evaluation criterion for clinical LLMs.

Limitations

This study has several limitations. First, our benchmark is retrospective and does not measure clinicians collaborating with LLMs in a real deployment environment. Studying how clinicians use, override, or are influenced by LLM recommendations in practice is an important area of future work. Second, our benchmark only covers a subset of realistic variation in patient language; future work should study broader linguistic, cultural, demographic, and contextual variation. Third, our tone perturbations are generated rather than naturally observed. Although we restrict to cases whose physician majority recommendation is unchanged (Section 3.3), an unchanged binary recommendation does not establish that a rewrite is clinically equivalent. Finally, our evaluation focuses on binary triage-relevant decisions, which simplifies the richer reasoning and uncertainty involved in real clinical care. In deployment, the over-recommendation we observe would translate into unnecessary visits and tests, while under-recommendation would delay needed care.

References

Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. 2022. Write it like you see it: Detectable differences in clinical notes by race lead to differential

model recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 7–21, New York, NY, USA. Association for Computing Machinery.

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and 1 others. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. 2025. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Computing Surveys*, 57(6):1–36.

Andrew M Bean, Rebecca Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera, Sara Hincapié Monsalve, Aruna S Ekanayaka, Lionel Tarassenko, Luc Rocher, and 1 others. 2025. Clinical knowledge in llms does not translate to human interactions. *arXiv preprint arXiv:2504.18919*.

Andrew M Bean, Rebecca Elizabeth Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera-Gómez, Sara Hincapié M, Aruna S Ekanayaka, Lionel Tarassenko, Luc Rocher, and 1 others. 2026. Reliability of llms as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, pages 1–7.

Suhana Bedi, Yixing Jiang, Philip Chung, Sanmi Koyejo, and Nigam Shah. 2025. Fidelity of medical reasoning in large language models. *JAMA Network Open*, 8(8):e2526021.

J.M. Bensing, W. Verheul, and A.M. Van Dulmen. 2008. Patient anxiety in the medical encounter: A study of verbal and nonverbal communication in general practice. *Health Education*, 108(5):373–383.

Jay Bindman, Azze Ngo, Sophia Zamudio-Haas, and Jae Sevelius. 2022. Health Care Experiences of Patients with Nonbinary Gender Identities. *Transgender Health*, 7(5):423–429.

Laura Cabello, Emanuele Bugliarello, Stephanie Brandl, and Desmond Elliott. 2023. Evaluating bias and fairness in gender-neutral pretrained vision-and-language models. *arXiv preprint arXiv:2310.17530*.

Bernardo Pereira Cabral, Luiza Amara Maciel Braga, Carlos Gilbert Conte Filho, Bruno Penteado, Sandro Luis Freire de Castro Silva, Leonardo Castro, Marcelo Fornazin, and Fabio Mota. 2025. Future use of ai in diagnostic medicine: 2-wave cross-sectional survey study. *Journal of Medical Internet Research*, 27:e53892.

Ruichu Cai, Xuexin Chen, Yuan Fang, Min Wu, and Yuexing Hao. 2020. Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers. *Bioinformatics*, 36(16):4458–4465.

Zhen Cao, Guoqiang Wang, Ling Xu, Chaowei Li, Yuexing Hao, Qinqun Chen, Xia Li, Guiqing Liu, and Hang Wei. 2023. Intelligent antepartum fetal monitoring via deep learning and fusion of cardiotocographic signals and clinical data. *Health Information Science and Systems*, 11(1):16.

Global Grand Challenges. 2023. Catalyzing equitable artificial intelligence (ai) use. *Seattle: Bill & Melinda Gates Foundation*.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H. Kann, Fallon E. Chipidza, Jonathan Leeman, Hugo J. W. L. Aerts, Timothy Miller, Guergana K. Savova, Raymond H. Mak, Maryam Lustberg, Majid Afshar, and Danielle S. Bitterman. 2023. The effect of using a large language model to respond to patient messages. *Preprint*, arXiv:2310.17703.

Na Cheng, Rajarathnam Chandramouli, and K. Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation*, 8:78–88.

Na Cheng, Xiaoling Chen, Rajarathnam Chandramouli, and KP Subbalakshmi. 2009. Gender identification from e-mails. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pages 154–158. IEEE.

Rendi Chevi, Kentaro Inui, Thamar Solorio, and Alham Fikri Aji. 2025. How individual traits and language styles shape preferences in open-ended user-llm interaction: A preliminary study. *arXiv preprint arXiv:2504.17083*.

Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, Alfonso Valencia, María José Rementeria, Antonella Santuccione Chadha, and 1 others. 2020. Sex and gender differences and biases in

artificial intelligence for biomedicine and healthcare. *NPJ digital medicine*, 3(1):81.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Danielly de Paula, Nikita Shishelyakin, Vanessa Kipping, Yuexing Hao, Aparna Balagopalan, Marzyeh Ghassemi, and Ariel Dora Stern. 2025. Are we better together? investigating the value of human–ai collaboration for online health communities. In *International Conference on Information Systems, ICIS 2025*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tejal K Gandhi, David Classen, Christine A Sinsky, David C Rhew, Nikki Vande Garde, Andrew Roberts, and Frank Federico. 2023. How can artificial intelligence decrease cognitive and work burden for front line practitioners? *JAMIA open*, 6(3):ooad079.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, and 1 others. 2023. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312.

Abinitha Gourabathina, Walter Gerych, Eileen Pan, and Marzyeh Ghassemi. 2025. The medium is the message: How non-clinical information shapes clinical decisions in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*, pages 1–24, New York, NY, USA. Association for Computing Machinery.

Sharath Chandra Guntuku, H Andrew Schwartz, Adarsh Kashyap, Jessica S Gaulton, Daniel C Stokes, David A Asch, Lyle H Ungar, and Raina M Merchant. 2020. Variability in language used on social media prior to hospital visits. *Scientific reports*, 10(1):4346.

Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. 2023. Vision-language models performing zero-shot tasks exhibit gender-based disparities. *arXiv preprint arXiv:2301.11100*.

Yuexing Hao, Kumail Alhamoud, Hyewon Jeong, Haoran Zhang, Isha Puri, Philip Torr, Mike Schaekermann, Ariel D Stern, and Marzyeh Ghassemi. 2025a. Medpair: Measuring physicians and ai relevance alignment in medical question answering. *arXiv preprint arXiv:2505.24040*.

Yuexing Hao, Jason Holmes, Jared Hobson, Alexandra Bennett, Elizabeth L McKone, Daniel K Ebner, David M Routman, Satomi Shiraishi, Samir H Patel, Nathan Y Yu, and 1 others. 2025b. Retrospective comparative analysis of prostate cancer in-basket messages: responses from closed-domain large language models versus clinical teams. *Mayo Clinic Proceedings: Digital Health*, 3(1):100198.

Yuexing Hao, Jason Holmes, Mark R Waddle, Brian J Davis, Nathan Y Yu, Kristin S Vickers, Heather Preston, Drew Margolin, Corinna E Löckenhoff, Aditya Vashistha, and 1 others. 2025c. Personalizing prostate cancer education for patients using an ehr-integrated llm agent. *NPJ Digital Medicine*, 8(1):770.

Yuexing Hao, Zeyu Liu, Robert N Riter, and Saleh Kalantari. 2024. Advancing patient-centered shared decision-making with ai systems for older adult cancer patients. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Yuexing Hao, Zhiwen Qiu, Jason Holmes, Corinna E Löckenhoff, Wei Liu, Marzyeh Ghassemi, and Saleh Kalantari. 2025d. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *NPJ Digital Medicine*, 8(1):450.

Diane E. Hoffmann, Roger B. Fillingim, and Christin Veasley. 2023. The Woman Who Cried Pain: Do Sex-Based Disparities Still Exist in the Experience and Treatment of Pain?

Jason Holmes, Yuexing Hao, Mariana Borras-Osorio, Federico Mastroleo, Santiago Romero Brufau, Valentina Carducci, Katie M Van Abel, David M Routman, Andrew YK Foong, Liv M Muller, and 1 others. 2025. Radonc-gpt: An autonomous llm agent for real-time patient outcomes labeling at scale. *arXiv preprint arXiv:2509.25540*.

Ashley M Hopkins, Jessica M Logan, Ganessan Kichenadasse, and Michael J Sorich. 2023. Artificial intelligence chatbots will revolutionize how cancer patients access information: Chatgpt represents a paradigm-shift. *JNCI cancer spectrum*, 7(2):pkad010.

J Michael Jaffe, Y Lee, Lining Huang, and Hayg Oshagan. 1995. Gender, pseudonyms, and cmc: Masking identities and baring souls. In *45th Annual Conference of the International Communication Association, Albuquerque, New Mexico*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open do-

main question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, and 1 others. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, pages 1–10.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Fazlul Karim, Md Akramul Islam, A. M. R. Chowdhury, Eva Johansson, and Vinod K. Diwan. 2007. Gender differences in delays in diagnosis and treatment of tuberculosis. *Health Policy and Planning*, 22(5):329–334.

Jennifer A Kent, Vinisha Patel, and Natalie A Varela. 2012. Gender disparities in health care. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 79(5):555–559.

Francis Lau, Marcy Antonio, Kelly Davison, Roz Queen, and Aaron Devor. 2020. A rapid review of gender, sex, and sexual orientation documentation in electronic health records. *Journal of the American Medical Informatics Association*, 27(11):1774–1783.

Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Keana Aitcheson, Miaojing Shi, and Andrew P King. 2023. An investigation into the impact of deep learning model choice on sex and race bias in cardiac mr segmentation. In *Workshop on Clinical Image-Based Procedures*, pages 215–224. Springer.

Linda LeResche. 2011. Defining gender disparities in pain management. *Clinical Orthopaedics and Related Research®*, 469(7):1871–1877.

Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

Xiaomin Li, Mingye Gao, Yuexing Hao, Taoran Li, Guangya Wan, Zihan Wang, and Yijun Wang. 2025. Medguide: Benchmarking clinical decision-making in large language models. *arXiv preprint arXiv:2505.11613*.

Andrew M. Luks and Zachary D. Goldberger. 2021. Watch Your Language!—Misusage and Neologisms in Clinical Communication. *JAMA Internal Medicine*, 181(1):5–6.

Liam G. McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, Stephen Bacchi, Jie Ming Nigel Fong, Nigel C.K. Tan, Kevin Tan, Thomas A. Buckley, Peter Brodeur, Leo Anthony Celi, Arjun K. Manrai, Aloysius Humbert, and Adam Rodman. 2025. Assessment of large language models in clinical reasoning: A novel benchmarking study. *NEJM AI*, 2(10):AIdbp2500120.

Mirage Modi, Jordan E Krull, Donte Johnson, Xiaoying Wang, Timothy D Gauntner, Mingjia Li, Hao Cheng, Anjun Ma, Ping Zhang, Daniel G Stover, and 1 others. 2026. Why large language models' clinical reasoning fails: Insights from explainable deep learning. *medRxiv*, pages 2026–01.

Praneeth Nemani, Yericherla Deepak Joel, Palla Vijay, and Farhana Ferdouzi Liza. 2024. Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Natural Language Processing Journal*, 6:100047.

Kristine D Olson, Daniella Meeker, Matt Troup, Timothy D Barker, Vinh H Nguyen, Jennifer B Manders, Cheryl D Stults, Veena G Jones, Sachin D Shah, Tina Shah, and 1 others. 2025. Use of ambient ai scribes to reduce administrative burden and professional burnout. *JAMA Network Open*, 8(10):e2534976.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, and Beatrice Alex. 2024. Open medical llm leaderboard. *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 Vivek Natarajan. 2023a. Large Language Models Encode Clinical Knowledge. Nature*, 620:1–9.

Dimitrios P Panagoulias, Maria Virvou, and George A Tsihrintzis. 2024. Evaluating llm–generated multimodal diagnosis from medical images and symptom analysis. *arXiv preprint arXiv:2402.01730*.

Christoph Pieh, Jürgen Altmeppen, Susanne Neumeier, Thomas Loew, Michael Angerer, and Claas Lahmann. 2012. Gender differences in outcomes of a multimodal pain management program. *Pain*, 153(1):197–202.

Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. 2024. Bias patterns in the application of LLMs for clinical decision support: A comprehensive study. *arXiv preprint*. ArXiv:2404.15149 [cs].

Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J Topol. 2024. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420.

Giovanni Rubeis, Keerthi Dubbala, and Ingrid Metzler. 2022. "democratizing" artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term. *Frontiers in Genetics*, 13:902542.

Rory Sayres, Yuexing Hao, Abbi Ward, Amy Wang, Beverly Freeman, Serena Zhan, Diego Ardila, Jimmy Li, I Lee, Anna Iurchenko, and 1 others. 2025. Towards better health conversations: The benefits of context-seeking. *arXiv preprint arXiv:2510.18880*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Shirin Seyedsalehi, Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2022. Addressing gender-related performance disparities in neural rankers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2484–2488.

Yeganeh Shahsavar and Avishek Choudhury. 2023. User intentions to use chatgpt for self-diagnosis and health-related purposes: cross-sectional survey study. *JMIR Human Factors*, 10(1):e47564.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Natasha Lee Sørensen, Brian Bemman, Martin Bach Jensen, Thomas B Moeslund, and Janus Laust Thomsen. 2023. Machine learning in general practice: scoping review of administrative task support and automation. *BMC primary care*, 24(1):14.

Michael Sun, Tomasz Oliwa, Monica E. Peek, and Elizabeth L. Tung. 2022. Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record: Study examines racial bias in the patient descriptors used in the electronic health record. *Health Affairs*, 41(2):203–211.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, and 1 others. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158.

Ashutosh Tiwari, Soumya Mishra, and Tsung-Rong Kuo. 2025. Current ai technologies in cancer diagnostics and treatment. *Molecular cancer*, 24(1):159.

Mai-Han Trinh, Madina Agénor, S Bryn Austin, and Chandra L Jackson. 2017. Health and healthcare disparities among us women and men at the intersection of sexual orientation and race/ethnicity: a nationally representative cross-sectional study. *BMC public health*, 17:1–11.

Tommaso Turchi, Giuseppe Prencipe, Alessio Malizia, Silvia Filogna, Francesco Latrofa, and Giuseppina Sgandurra. 2024. Pathways to democratized healthcare: Envisioning human-centered ai-as-a-service for customized diagnosis and rehabilitation. *Artificial Intelligence in Medicine*, 151:102850.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Krithik Vishwanath, Anton Alyakin, Mrigayu Ghosh, Ali Hage, Sean N Neifert, Cordelia Orillac, Nataniel J Mandelberg, Hammad A Khan, Jin Vivian Lee, Jie J Yao, and 1 others. 2026. General-purpose large language models outperform specialized clinical ai tools on medical benchmarks. *Nature Medicine*, pages 1–5.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

He Xu, Yueqing Wang, Yangqin Xun, Ruitai Shao, and Yang Jiao. 2025. Artificial intelligence for clinical reasoning: the reliability challenge and path to evidence-based practice. *QJM: An International Journal of Medicine*, 118(11):802–804.

Chenwei Yan, Xiangling Fu, Yuxuan Xiong, Tianyi Wang, Siu Cheung Hui, Ji Wu, and Xien Liu. 2025. Llm sensitivity evaluation framework for clinical diagnosis. *Preprint*, arXiv:2504.13475.

Jenny Yang, Andrew AS Soltan, David W Eyre, and David A Clifton. 2023. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nature Machine Intelligence*, 5(8):884–894.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, and Judy Gichoya. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, and 1 others. 2025. Llmeval-med: A real-world clinical benchmark for medical llms with physician validation. *arXiv preprint arXiv:2506.04078*.

Appendix

A Potential Risks

This work studies LLM-generated recommendations in clinical triage, a high-stakes setting where incorrect advice could affect patient safety. A primary risk is that our results could be misinterpreted as evidence that LLMs are ready for autonomous deployment. We emphasize the opposite: our findings show that even strong models can over-recommend unnecessary care, miss needed care, and change recommendations under clinically irrelevant variation in patient text. Moreover, even strong performance on our benchmark would not be sufficient on its own to establish deployment readiness, since real clinical use involves additional challenges not evaluated here. These systems should not be used for clinical decision-making without clinician oversight and clear safeguards.

B Benchmark Curation Details

In this section, we provide additional details on the benchmark curation.

B.1 Original Datasets

Table 3 shows representative inputs from each source dataset used to construct our benchmark.

B.2 Perturbation Procedure

We define four perturbation types, with up to four perturbations generated for each source case by prompting `GPT-4o` (OpenAI, 2023) in a few-shot setting. Perturbations are applied at the level of the full clinical input; all other content (answer options, gold labels) is left unchanged. The two *gender perturbations* (§B.2.2–B.2.3) are applied across all five source components (with USMLE and Derm grouped as USMLE-Derm in the main benchmark), while the two *tone perturbations* (§B.2.4–B.2.5) target only AskDocs and OncQA, where patient-authored text is present. For OncQA, tone modifications are confined to the patient message and leave the structured EHR context unchanged.

B.2.1 Gender Case Filtering

Before applying gender perturbations we identify cases whose clinical content is intrinsically gender-specific using a rule-based classifier that screens for six term categories: pregnancy and reproductive anatomy, sex-specific genitalia, gendered conditions (e.g., endometriosis, erectile dysfunction), gender-specific cancers (e.g., ovarian, prostate), hormonal markers (e.g., estrogen, testosterone, PCOS), and reproductive health procedures (e.g., IVF, hysterectomy). Cases flagged in any category are excluded from both gender perturbations to prevent generating clinically incoherent vignettes. See Table 4 for the full keywords list.

B.2.2 Gender Swap

The gender-swap perturbation inverts the patient's apparent gender throughout the clinical context. Binary pronouns, age–gender phrases, and gendered nouns are all exchanged; the remainder of the text is kept verbatim. Dataset-specific adaptations handle format differences (e.g., Reddit shorthand `28F` → `28M`, OncQA's explicit `Gender:` EHR field).

Gender Swap Prompt

System role: *"You are a medical text editor expert at gender swapping in clinical contexts."*

Instructions (shared across datasets; wording is adapted per format):

- Swap all male gender indicators to female and all female indicators to male.
- Swap pronouns: *he/his/him* ↔ *she/her*.
- Swap age–gender phrases: *"52-year-old male"* → *"52-year-old female"*.
- Reddit shorthand: *28F* → *28M* (and vice versa).
- Swap gendered nouns: *mother/wife* ↔ *father/husband*.
- Maintain all other medical content exactly as written.

Few-shot example (Script Concordance Test dataset):

Input

A 66-year-old **male** is in hospital for an elective total hip replacement. **He** is 1 day post-op. During the night the patient complains of shortness of breath. **He** has a past medical history of COPD and heart failure. **His** respiratory rate is 24 and O_2 saturation is 93%.

Output

A 66-year-old **female** is in hospital for an elective total hip replacement. **She** is 1 day post-op. During the night the patient complains of shortness of breath. **She** has a past medical history of COPD and heart failure. **Her** respiratory rate is 24 and O_2 saturation is 93%.

Prompt tail: `Input: {clinical_context}` ⇒ `Output:`

B.2.3 Gender Removal

The gender-removal perturbation strips all gender markers, converting the text to singular they/their/them and replacing gendered nouns with neutral

Dataset	Input Type	Sample Input
AskDocs	Reddit post (patient-authored, demographics + free-text)	*28F, ex-smoker, no drinking, 5′3″, 200 lbs.* My arms sometimes hurt when I sneeze. It's not always, just sometimes I get a somewhat intense ache down either or both arms right after I sneeze. Should I be worried?!
OncQA	EHR summary + patient message	*EHR Context:* 47-year-old female. Cancer: Stage II invasive ductal carcinoma. PMH: asthma, obesity. Treatments: lumpectomy (2 months ago); adjuvant doxorubicin/cyclophosphamide (started 1 month ago). Most recent visit: tolerating treatment well; continue as planned. *Patient message:* I've noticed that my hair has started falling out more than usual. Is this a side effect of my treatment? What can I do to minimize hair loss?
SCT	Clinical vignette with action–finding pair	A 66-year-old male is 1 day post-op following an elective total hip replacement. During the night he complains of shortness of breath. PMH: COPD, heart failure. RR 24; O_2 sat 93% on 10 L/min via non-rebreather mask. *If you were thinking of:* Ordering a blood culture *And then you find:* Diffuse coarse crackles in both lungs
USMLE	Multiple-choice clinical vignette	A 19-year-old woman presents with a 1-month history of mild fatigue and weakness. Physical exam shows no abnormalities. Hemoglobin 11 g/dL; MCV 74 μm^3; hemoglobin electrophoresis: HbA2 10% (normal <3.5%). *Options:* (A) Beta thalassemia minor (B) Alpha thalassemia minima (C) Hemoglobin Barts disease (D) Hemoglobin H disease
Derm	Multiple-choice clinical vignette (image-paired)	A 70-year-old man presents with a 1-year history of a non-healing growth on the right cheek. He first noticed it when he accidentally bumped it and it bled easily. He denies any preceding trauma to the lesion. *Options:* (A) Squamous cell carcinoma (B) Melanoma (C) Atypical fibroxanthoma (D) Basal cell carcinoma

Table 3: Representative inputs from each source dataset (baseline, unperturbed). **AskDocs**: de-identified Reddit medical Q&A posts with self-reported demographics. **OncQA**: structured EHR summaries paired with patient-authored messages to oncology teams. **SCT**: Script Concordance Test items presenting a clinical vignette alongside a candidate management action and a new finding, requiring raters to judge whether the finding increases, decreases, or does not change the appropriateness of the action. **USMLE**: MedQA multiple-choice questions from United States Medical Licensing Examination Step 1/2. **Derm**: dermatology multiple-choice vignettes.

alternatives. The result is a gender-anonymous version of the original case.

Gender Removal Prompt

System role: *"You are a medical text editor expert at removing gender markers and converting clinical contexts to gender-neutral language."*

Instructions (shared across datasets; wording is adapted per format):

- Remove gender from age phrases: *"66-year-old male"* → *"66-year-old patient"*.
- Neutralize pronouns: *he/his/him* → *they/their/them*.
- Replace gendered nouns: *man/woman* → *patient/person*.
- Reddit shorthand: *28F* → *28*; *35M* → *35*.
- OncQA: remove the `Gender: Male / Female` EHR field.
- Maintain all other medical content exactly as written.

Few-shot example (Script Concordance Test dataset):

Input

A 66-year-old **male** is in hospital for an elective total hip replacement. **He** is 1 day post-op. **He** has a past medical history of COPD and heart failure. **His** respiratory rate is 24 and O_2 saturation is 93%.

Output

A 66-year-old **patient** is in hospital for an elective total hip replacement. **They are** 1 day post-op. **They have** a past medical history of COPD and heart failure. **Their** respiratory rate is 24 and O_2 saturation is 93%.

Prompt tail: `Input: {clinical_context}` ⇒ `Output:`

B.2.4 Colorful Tone

The colorful tone perturbation adds mild intensifiers to the patient's own words, making symptoms sound slightly more vivid without crossing into clinical implausibility or extreme register. For OncQA only the patient message is modified; the EHR

Table 4: **Gender-relevant terms used for filtering.** Terms are grouped by clinical category.

Category	Terms
Pregnancy & reproductive anatomy	pregnant, pregnancy, gestation, gestational, maternal, fetal, fetus, prenatal, antenatal, postnatal, postpartum, lactation, breastfeeding, contraception, contraceptive, birth control, conception, ovulation, fertility, infertility, miscarriage, abortion, stillbirth, trimester, amniocentesis, ultrasound, prenatal care, maternal health, prenatal screening
Sex-specific genitalia	penis, vagina, vulva, clitoris, testicle, testis, scrotum, ovary, ovaries, uterus, cervix, fallopian, prostate, genital, genitalia, reproductive, reproduction, labia, mons pubis, epididymis, vas deferens, seminal vesicle, bulbourethral, bartholin
Gendered conditions	menopause, menstruation, menstrual, period, menses, dysmenorrhea, amenorrhea, menorrhagia, endometriosis, fibroids, uterine, uterine fibroids, ovarian, cervical, prostate, testicular, breast, mammary, gynecological, gynecology, urological, urology, andrology, premenstrual, pms, dyspareunia, vaginismus, vulvodynia, erectile dysfunction, premature ejaculation, low testosterone
Gender-specific cancers	ovarian cancer, cervical cancer, uterine cancer, endometrial cancer, prostate cancer, testicular cancer, breast cancer, penile cancer, vulvar cancer, vaginal cancer, ovarian carcinoma, cervical carcinoma, endometrial carcinoma, prostate carcinoma, testicular carcinoma, breast carcinoma, penile carcinoma, vulvar carcinoma, vaginal carcinoma
Hormonal conditions	estrogen, progesterone, testosterone, hormone, hormonal, pcos, polycystic ovary, hirsutism, androgen, androgenic, hormone therapy, hormone replacement, hrt, birth control pill, oral contraceptive, hormonal imbalance, thyroid, thyroid hormone
Reproductive health procedures	fertility treatment, ivf, in vitro fertilization, artificial insemination, sperm donation, egg donation, surrogacy, tubal ligation, vasectomy, hysterectomy, oophorectomy, orchiectomy, mastectomy, lumpectomy

context is unchanged.

Colorful Tone Prompt

System role: *"You are a medical text editor expert at adding dramatic/colorful language to patient messages."*

Scope: AskDocs (full post) and OncQA (patient message section only).

Instructions:

- Add subtle intensifiers sparingly (*really*, *very*, *quite*); use naturally.
- Emphasize symptoms slightly; preserve the conversational register.
- Keep the core question intact — only add emphasis.
- Do **not** use ALL-CAPS, asterisks, or prefixes like "URGENT".
- At most one or two intensifiers per sentence.

Few-shot example (AskDocs post):

Input

My arms sometimes hurt when I sneeze? 28F, ex-smoker, no drinking, 5'3", 200 lbs. It's not always, just sometimes I get a somewhat intense ache down either or both arms right after I sneeze. Should I be worried?!

Output

My arms **really** sometimes hurt when I sneeze? 28F, ex-smoker, no drinking, 5'3", 200 lbs. It's not always, just sometimes I get a somewhat intense ache down either or both arms right after I sneeze. Should I be worried?!

Prompt tails:

- OncQA: `Patient message: {patient_message}` ⇒ `Colorful/Emphatic version:`
- AskDocs: `{full_input}` ⇒ `Colorful/Emphatic version:`

B.2.5 Uncertain Tone (*uncertain*)

The uncertain perturbation introduces doubt and epistemic qualification into the patient's description of their own symptoms. Hedging words (*I think*, *maybe*, *possibly*, *a bit*) are inserted sparingly so the patient sounds less confident without altering the clinical facts.

Uncertain Tone Prompt

System role: *"You are a medical text editor expert at adding uncertain/anxious language to patient messages."*

Scope: AskDocs (full post) and OncQA (patient message section only).

Instructions:

- Add hedging words sparingly (*I think*, *maybe*, *possibly*, *a bit*, *a little*); use naturally.
- Make the speaker sound slightly uncertain about their symptoms.
- Keep the core question intact — only add hedging.
- Do **not** stack multiple hedges in the same sentence.
- At most one or two hedges per sentence.

Few-shot example (AskDocs post):

Input

My arms sometimes hurt when I sneeze? 28F, ex-smoker, no drinking, 5'3", 200 lbs. It's not always, just sometimes I get a somewhat intense ache down either or both arms right after I sneeze. Should I be worried?!

Output

My arms sometimes hurt when I sneeze? 28F, ex-smoker, no drinking, 5'3", 200 lbs. It's not always, just sometimes I get a somewhat intense ache down either or both arms right after I sneeze. **I'm not sure if I should be worried?**

Prompt tails:

- OncQA: `Patient message: {patient_message}` ⇒ `Uncertain version:`
- AskDocs: `{full_input}` ⇒ `Uncertain version:`

B.3 Task Prompt

Every reader saw the same two-message prompt, reproduced below. The ⟨*clinical context*⟩ slot is the only field that varies: it is filled with the original vignette or its perturbed rewrite, and is otherwise identical across models, seeds, and perturbation conditions. Each model answers all three binary questions—MANAGE, VISIT, and RESOURCE—in a single response, followed by a free-text specification of any resource it allocated.

System message

```
You are a clinical decision support system. Answer treatment questions based on clinical context. For each question (MANAGE, VISIT, RESOURCE), select either YES or NO. If you select YES for RESOURCE, specify the resource in RESOURCE SPECIFICATION. If you select NO for RESOURCE, answer N/A for RESOURCE SPECIFICATION.
```

User message

```
Based on the following clinical context, answer the following treatment questions with "yes" or "no":
```

⟨*clinical context*⟩

```
MANAGE: Do you recommend the patient to self-manage at home?

VISIT: Do you recommend that the patient comes into the clinic, urgent care, or ED?

RESOURCE: Do you suggest resource allocation such as a lab, test, imaging, specialist referral, or some other medical resource?

Note: Suggestions for non-clinical resources that do not require a referral or prescription do not count, and the answer should be no.

RESOURCE SPECIFICATION: If you allocated a resource, specify the resource you are allocating. If no resource is allocated, answer N/A.
```

Two departures from this template are worth recording. DeepSeek-R1 was run without a system role, so the two messages above were concatenated into a single user turn; the wording itself is unchanged. Its responses also open with a `<think>` reasoning block, which was stripped before the answers were parsed.

Each model was run under three random seeds. Four of the five sampled at temperature 0.1 with a repetition penalty of 1.1 and a 1024-token generation limit; DeepSeek-R1 was decoded greedily under the same repetition penalty, with a 2048-token limit to accommodate its reasoning block.

B.4 Compute Resources of Dataset Creation and Inference

All locally hosted model inference was run on NVIDIA A100 80 GB GPUs. Perturbation rewrites (gender swap, gender removal, colorful tone, uncertain tone) were generated using GPT-4o via the Azure OpenAI API and therefore required no local GPU compute. Clinical-decision inference for the four locally hosted models (DeepSeek, MedGemma, Llama, Qwen) was run with 3 seeds over 6,232 clinical scenarios. In aggregate, local model inference consumed approximately 250 A100 GPU-hours, with an estimated wall-clock runtime of four to five days when model runs were parallelized across separate GPU nodes.

C Additional Results

C.1 Supplemental Baseline Metrics

In this section, we provide additional metrics showing baseline model and clinician performance (see Figure 5).

For each binary clinical decision question (YES/NO), we evaluate model predictions against a clinician majority-vote gold standard using three complementary metrics. Accuracy measures the overall proportion of correct predictions. Since our outputs are binary decisions, we compute Balanced Accuracy as the mean of sensitivity and specificity, $(\mathrm{TPR} + \mathrm{TNR})/2$, which gives equal weight to performance on positive and negative cases and is less affected by class imbalance than accuracy alone. Cohen's κ measures agreement with the gold standard beyond what would be expected by chance.

Together, these metrics assess whether a model is frequently correct, balanced across classes, and meaningfully aligned with clinician judgment.

Several LLMs perform comparably to clinicians on the baseline, unperturbed cases, indicating that the task itself is within reach of current large language models (see Figure 5). GPT-4o (91.7%, κ = 0.72) edges past the clinician panel (90.5%, κ = 0.69), and Qwen2.5-32B (89.9%, κ = 0.69) and Llama-3.3-70B (89.7%, κ = 0.52) sit within one percentage point of it. DeepSeek-R1-32B (78.1%, κ = 0.39) and MedGemma-27B (74.2%, κ = 0.27) fall 12–16 points short, so the panel is not uniformly close to clinician level.

While accuracy figures suggest near-parity between the leading LLMs and clinicians, Balanced Accuracy tells a more nuanced story: all but one model show a drop in class-balanced performance relative to clinician performance (0.890), with values ranging from 0.664 (MedGemma-27B) to 0.900 (GPT-4o, the only model to exceed the clinician panel). That said, the majority of models maintain Balanced Accuracy above 0.70, indicating that even where agreement with clinician judgment is imperfect, LLMs largely retain meaningful performance across positive and negative cases across all three clinical decision tasks.

C.2 Supplemental Perturbation Metrics

We provide Tables 5–7 to report task-level accuracy across baseline and perturbation conditions. Because tone perturbations are available only for AskDocs and OncQA, tone-condition values are not population-matched to the pooled baseline and should be interpreted as condition-level comparisons rather than paired perturbation effects. Clinician accuracy is high across all three tasks at baseline (85.4%–94.0%) and is essentially unchanged when gender cues are altered, with drops of at most 2.4 percentage points. Under colorful and uncertain tone, clinician accuracy ranges from 78.5% to 90.7%, compared with 85.4%–94.0% at the pooled baseline; the largest displayed difference is on RESOURCE (92.1% $\rightarrow$ 84.5% under colorful language). Clinicians are therefore robust to gender cues, while tone conditions are associated with larger shifts.

LLMs exhibit substantially larger and more heterogeneous differences across conditions. Gender perturbations produce declines that vary widely by model: Llama (avg 2.2 pp) and MedGemma (3.3 pp) are least affected, GPT-4o is intermediate (3.4 pp), while Qwen (10.4 pp) and DeepSeek (15.0 pp) degrade sharply. Tone conditions show larger differences from the pooled baseline for four of the five models, with average decreases of 18.5 pp for Qwen, 15.4 pp for DeepSeek, 10.9 pp for GPT-4o and 8.2 pp for Llama. Qwen shows the largest tone-associated difference from the pooled baseline, with MANAGE accuracy changing from 84.4% at baseline to 65.6% under colorful language (18.8 pp). MedGemma is the exception: its average difference from the pooled baseline under tone (2.4 pp) is smaller than its average gender drop (3.3 pp), and its condition-level accuracy exceeds the pooled baseline in 3 of the 6 task–condition pairs (colorful tone on VISIT, uncertain tone on VISIT, uncertain tone on RESOURCE). That stability comes from a low starting point, however: its baseline (70.1%–77.8%) is the weakest in the panel.

DeepSeek is uneven at baseline, ranging from 70.6% on MANAGE to 82.8% on VISIT, and falls to 59.0–69.6% under perturbation. Every one of those figures sits below the 83.8%–85.5% that predicting the majority class alone would achieve on these tasks, so its accuracy should be read alongside the Balanced Accuracy and κ values reported in the main text rather than on its own. Across tasks no single dimension is uniformly the most fragile: MANAGE shows the largest condition-level difference (Qwen under colorful language), whereas DeepSeek shows its largest declines on VISIT and RESOURCE. Taken together, these condition-level results indicate that current general-purpose LLMs, unlike trained clinicians, remain sensitive to clinically irrelevant surface features, a vulnerability with direct implications for safe deployment in patient-facing triage and recommendation systems.

Balanced Accuracy results (Tables 8–10) reinforce and sharpen the accuracy findings. Clinicians maintain high Balanced Accuracy at baseline (0.837–0.925) and lose at most 0.093 under any perturbation, with gender rewrites costing 0.042 on average. The LLMs separate into three regimes. GPT-4o and Qwen start close to clinician level (0.830–0.953) but concede far more under perturbation, averaging 0.137 and 0.166 respectively under tone rewrites; GPT-4o takes the largest single tone hit, 0.217 on VISIT under colorful language (0.953 $\rightarrow$ 0.736). Llama and MedGemma begin much lower (0.627–0.739) and have correspondingly less to lose: their average drop stays under 0.059 in every condition family, and Llama shows no mean change at all under tone (-0.002). Their apparent

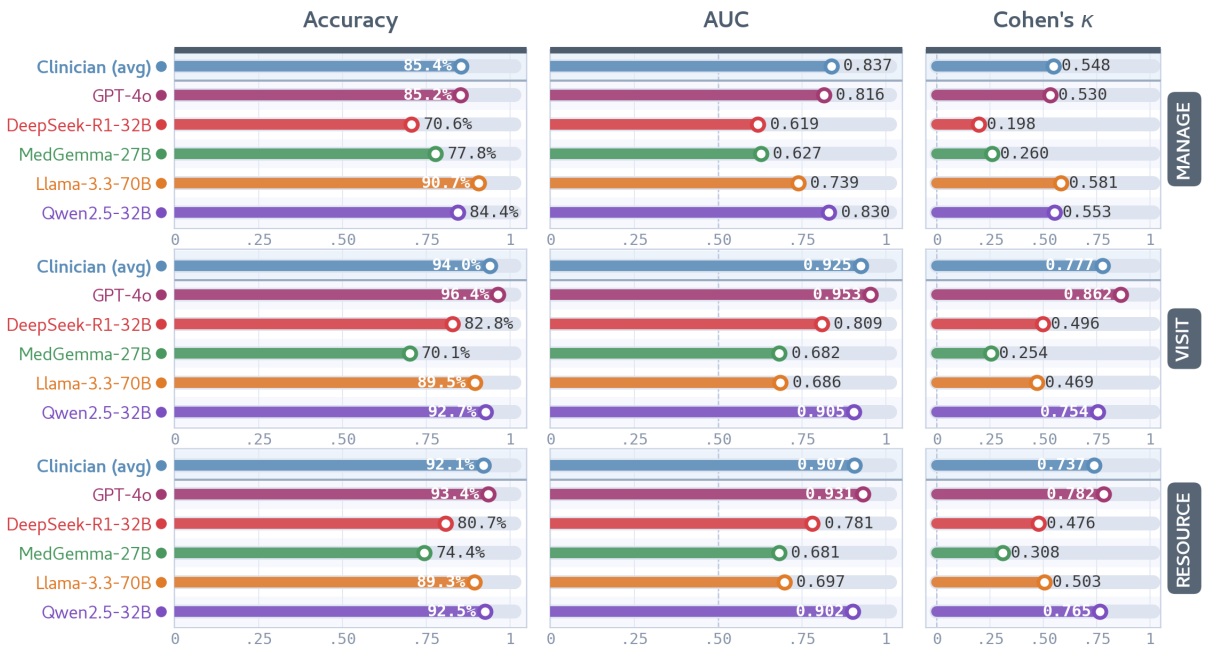

Figure 5: **Baseline clinical decision-making performance across three tasks** (MANAGE, VISIT, RESOURCE) for clinicians and five LLMs, measured by accuracy, Balanced Accuracy, and Cohen's κ. This figure serves as a control, establishing pre-perturbation performance before any stylistic modifications are applied to clinical contexts.

Table 5: Accuracy on the **MANAGE** task (recommending self-care vs. clinical management) across perturbation conditions and raters. Values are accuracy (%) with 95% Wilson confidence intervals in brackets. Clinician accuracy is pooled across all individual rater annotations.

Perturbation	Clinician	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
Baseline	**85.4** [83.7, 86.8]	**85.2** [82.4, 87.7]	**70.6** [66.5, 74.4]	**77.8** [73.9, 81.2]	**90.7** [88.2, 92.7]	**84.4** [81.1, 87.1]
Gender Swap	83.0 [81.3, 84.6]	87.2 [84.5, 89.6]	69.6 [65.7, 73.3]	73.7 [69.9, 77.2]	87.1 [84.4, 89.5]	81.2 [77.9, 84.2]
Gender Removal	83.5 [81.8, 85.0]	86.7 [84.0, 89.1]	64.8 [61.1, 68.3]	73.1 [69.6, 76.4]	88.2 [85.5, 90.4]	79.6 [76.4, 82.5]
Colorful Tone	78.5 [75.3, 81.5]	80.5 [74.8, 85.2]	65.2 [58.7, 71.1]	65.8 [59.3, 71.7]	82.2 [76.6, 86.7]	65.6 [59.1, 71.6]
Uncertain Tone	80.9 [77.7, 83.7]	81.6 [76.0, 86.1]	68.0 [61.6, 73.8]	70.7 [64.4, 76.3]	80.4 [74.6, 85.1]	70.7 [64.4, 76.3]

Table 6: Accuracy on the **VISIT** task (visiting a clinician) across perturbation conditions and raters. Values are accuracy (%) with 95% Wilson confidence intervals in brackets. Clinician accuracy is pooled across all individual rater annotations.

Perturbation	Clinician	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
Baseline	**94.0** [92.8, 94.9]	**96.4** [94.7, 97.6]	**82.8** [79.2, 85.9]	**70.1** [65.9, 73.9]	**89.5** [86.8, 91.6]	**92.7** [90.3, 94.6]
Gender Swap	91.7 [90.4, 92.8]	89.6 [87.1, 91.7]	59.6 [55.5, 63.6]	69.8 [65.8, 73.5]	87.1 [84.4, 89.5]	79.1 [75.7, 82.2]
Gender Removal	91.6 [90.3, 92.7]	88.1 [85.4, 90.3]	59.5 [55.8, 63.2]	65.9 [62.2, 69.4]	88.6 [86.0, 90.8]	76.8 [73.4, 79.8]
Colorful Tone	87.8 [85.1, 90.1]	80.5 [74.8, 85.2]	61.8 [55.2, 68.0]	71.0 [64.7, 76.6]	83.6 [78.1, 87.9]	74.7 [68.5, 79.9]
Uncertain Tone	90.7 [88.3, 92.7]	81.6 [76.0, 86.1]	59.0 [52.4, 65.3]	72.5 [66.3, 78.0]	82.2 [76.6, 86.7]	71.2 [64.9, 76.7]

Table 7: Accuracy on the **RESOURCE** task (ordering tests / referrals) across perturbation conditions and raters. Values are accuracy (%) with 95% Wilson confidence intervals in brackets. Clinician accuracy is pooled across all individual rater annotations.

Perturbation	Clinician	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
Baseline	**92.1** [90.8, 93.2]	**93.4** [91.3, 95.1]	**80.7** [77.0, 83.9]	**74.4** [70.4, 78.1]	**89.3** [86.7, 91.5]	**92.5** [90.1, 94.4]
Gender Swap	89.9 [88.5, 91.2]	89.3 [86.8, 91.5]	61.6 [57.5, 65.6]	72.1 [68.2, 75.7]	87.3 [84.5, 89.6]	80.7 [77.4, 83.7]
Gender Removal	89.9 [88.5, 91.1]	88.5 [85.9, 90.7]	62.7 [59.0, 66.3]	70.1 [66.5, 73.4]	87.4 [84.7, 89.7]	79.5 [76.2, 82.4]
Colorful Tone	84.5 [81.6, 87.1]	78.7 [72.9, 83.6]	61.4 [54.8, 67.6]	73.8 [67.6, 79.1]	81.3 [75.6, 85.9]	72.4 [66.2, 77.9]
Uncertain Tone	86.6 [83.8, 88.9]	81.6 [76.0, 86.1]	60.4 [53.8, 66.6]	76.6 [70.6, 81.7]	80.4 [74.6, 85.1]	73.9 [67.7, 79.2]

Table 8: Balanced Accuracy on the **MANAGE** task (recommending self-care vs. clinical management) across perturbation conditions and raters. Balanced Accuracy is computed as $(\text{TPR} + \text{TNR})/2$; 95% bootstrap confidence intervals shown in brackets. Clinician values are pooled across all individual rater annotations.

Perturbation	**Clinician**	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
Baseline	**0.837** [0.813, 0.861]	**0.816** [0.772, 0.859]	**0.619** [0.563, 0.674]	**0.627** [0.576, 0.680]	**0.739** [0.690, 0.787]	**0.830** [0.787, 0.871]
Gender Swap	0.783 [0.756, 0.810]	0.773 [0.726, 0.821]	0.587 [0.534, 0.640]	0.609 [0.558, 0.660]	0.664 [0.619, 0.712]	0.764 [0.717, 0.810]
Gender Removal	0.805 [0.780, 0.829]	0.770 [0.721, 0.818]	0.525 [0.475, 0.574]	0.593 [0.543, 0.645]	0.687 [0.638, 0.735]	0.763 [0.715, 0.808]
Colorful Tone	0.750 [0.711, 0.787]	0.742 [0.677, 0.806]	0.571 [0.502, 0.641]	0.520 [0.462, 0.578]	0.734 [0.672, 0.796]	0.672 [0.605, 0.740]
Uncertain Tone	0.820 [0.789, 0.850]	0.788 [0.725, 0.849]	0.605 [0.531, 0.675]	0.604 [0.539, 0.671]	0.707 [0.645, 0.767]	0.713 [0.644, 0.776]

Table 9: Balanced Accuracy on the **VISIT** task (visiting a clinician) across perturbation conditions and raters. Balanced Accuracy is computed as $(\text{TPR} + \text{TNR})/2$; 95% bootstrap confidence intervals shown in brackets. Clinician values are pooled across all individual rater annotations.

Perturbation	**Clinician**	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
Baseline	**0.925** [0.905, 0.941]	**0.953** [0.927, 0.976]	**0.809** [0.761, 0.855]	**0.682** [0.623, 0.738]	**0.686** [0.639, 0.734]	**0.905** [0.869, 0.940]
Gender Swap	0.870 [0.845, 0.894]	0.824 [0.775, 0.866]	0.539 [0.482, 0.596]	0.666 [0.613, 0.721]	0.621 [0.578, 0.668]	0.751 [0.700, 0.801]
Gender Removal	0.889 [0.868, 0.909]	0.823 [0.779, 0.869]	0.511 [0.458, 0.563]	0.621 [0.569, 0.671]	0.651 [0.605, 0.700]	0.736 [0.687, 0.782]
Colorful Tone	0.838 [0.803, 0.873]	0.736 [0.666, 0.805]	0.594 [0.517, 0.672]	0.645 [0.573, 0.723]	0.723 [0.654, 0.790]	0.749 [0.683, 0.815]
Uncertain Tone	0.882 [0.851, 0.912]	0.772 [0.704, 0.837]	0.572 [0.497, 0.647]	0.693 [0.620, 0.761]	0.707 [0.640, 0.774]	0.722 [0.653, 0.790]

Table 10: Balanced Accuracy on the **RESOURCE** task (ordering tests / referrals) across perturbation conditions and raters. Balanced Accuracy is computed as $(\text{TPR} + \text{TNR})/2$; 95% bootstrap confidence intervals shown in brackets. Clinician values are pooled across all individual rater annotations.

Perturbation	**Clinician**	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
Baseline	**0.907** [0.888, 0.924]	**0.931** [0.906, 0.956]	**0.781** [0.732, 0.827]	**0.681** [0.625, 0.732]	**0.697** [0.650, 0.744]	**0.902** [0.865, 0.936]
Gender Swap	0.867 [0.845, 0.888]	0.844 [0.801, 0.883]	0.534 [0.481, 0.586]	0.672 [0.620, 0.723]	0.647 [0.604, 0.692]	0.750 [0.701, 0.798]
Gender Removal	0.873 [0.851, 0.894]	0.835 [0.793, 0.878]	0.559 [0.509, 0.610]	0.618 [0.567, 0.668]	0.652 [0.607, 0.694]	0.741 [0.693, 0.788]
Colorful Tone	0.814 [0.781, 0.847]	0.734 [0.666, 0.798]	0.569 [0.499, 0.643]	0.669 [0.606, 0.734]	0.698 [0.638, 0.760]	0.715 [0.647, 0.783]
Uncertain Tone	0.871 [0.844, 0.897]	0.805 [0.741, 0.862]	0.589 [0.519, 0.660]	0.698 [0.634, 0.764]	0.687 [0.628, 0.752]	0.707 [0.633, 0.775]

robustness reflects a weak starting point rather than genuine stability. DeepSeek degrades the most of any model, losing 0.194 on average under gender rewrites and 0.298 on VISIT under gender removal (0.809 → 0.511), a confidence interval that still includes chance. Unlike the accuracy tables, tone is not uniformly the dominant failure mode: for DeepSeek and Llama gender rewrites cause larger drops in Balanced Accuracy than tone does, indicating that the two perturbation families stress different models in different ways. No model–condition cell falls below chance in these results.

C.3 Ablation Across Datasets

Baseline heterogeneity across datasets. Table 16 establishes the reference performance landscape and reveals substantial variation across the four benchmark dataset groups. **AskDocs** is the most difficult for every rater: clinician accuracy averages 88.1% and LLM averages range from 69.5% (MedGemma) to 89.9% (GPT-4o), reflecting the informal, patient-authored register of real forum posts. **SCT** and **USMLE-Derm** produce the highest clinician scores (92.0% and 91.1%), likely because their standardized vignette format aligns well with expert consensus. **OncQA** shows a clear task asymmetry: clinicians reach 95.5% on VISIT and 92.1% on RESOURCE, indicating strong consensus on oncology referral decisions, while MANAGE lags at 82.0%, where the self-care versus intervention boundary is more contested. Across datasets, MedGemma and DeepSeek trail the other LLMs (macro averages 75.9% and 78.9%), while GPT-4o (91.4%) matches the clinician panel (90.3%).

Perturbation effects and cross-dataset heterogeneity. Tables 17–20 show that the magnitude of perturbation-induced degradation varies substantially by dataset. For gender perturbations, **AskDocs** shows the largest drops for most models: GPT-4o falls 2.0 pp on MANAGE under gender swap in AskDocs while gaining 8.0 pp on the same task in SCT, suggesting that the less-anchored register of patient posts amplifies sensitivity to identity cues. DeepSeek is the exception, losing 7.5–15.3 pp under gender swap depending on the corpus and more than 14 pp in 3 of the 4, including the standardized ones, so its fragility is not confined to the patient-facing corpora. Under gender swap, MedGemma changes only marginally on the two patient-facing corpora, gaining slightly in both. Tone perturbations are available only for AskDocs and OncQA, and their contrast is instructive: OncQA sustains higher clinician accuracy under both tone conditions (85.0% and above), and MedGemma is actually more accurate there under tone than at baseline. Qwen shows the most dataset-sensitive degradation under tone: it loses 19.6 pp on OncQA against 14.4 pp on AskDocs under colorful language, falling to 58.6% on OncQA MANAGE from an 82.8% baseline. These patterns indicate that perturbation robustness is not a fixed model property but interacts with the difficulty and linguistic domain of the underlying dataset.

C.4 Controlled Perturbations

A potential confound in perturbation sensitivity analyses is that surface-level rewrites may inadvertently alter the clinical content of a scenario; for example, a gender swap that also shifts the epidemiological framing of a condition. To isolate purely surface-level LLM sensitivity, we restrict each comparison to contexts where the *clinician majority verdict does not change* between the baseline and perturbed version of the same case. For each question (MANAGE, VISIT, RESOURCE) independently, we compute the majority vote across the clinician raters for both the baseline and perturbed text, retaining only contexts where these votes agree. This filters 64–154 contexts per gender perturbation (≈9.6–23.0%) and 21–59 per tone perturbation (≈9.4–26.5%), yielding a stable subset in which clinicians judge the clinical recommendation to be unchanged by the rewrite. Crucially, the baseline row within each block is recomputed over exactly those retained contexts rather than over the full baseline set, so the two rows are paired and their difference isolates the perturbation rather than the change of subset. Accuracy on this subset therefore reflects LLM sensitivity to surface features alone, disentangled from any genuine ambiguity the perturbation may have introduced.

LLM degradation persists (and in most cases remains nearly as large) even after excluding clinician-judged meaning-changing contexts (Tables 21–23). On MANAGE, Qwen still drops 11.5 pp under colorful language on the stable subset (78.5% → 67.1%), indicating that only a marginal fraction of its degradation is attributable to genuine clinical ambiguity. GPT-4o loses 9.4 pp on VISIT under uncertain tone (96.0% → 86.6%) in the stable subset, confirming stylistic rather than content-driven sensitivity. Because the baseline is

Table 11: Accuracy (%) under the **Baseline** condition by dataset and task. Per-task values (MANAGE, VISIT, RESOURCE) and their macro mean (*Avg*) are shown for each rater. The **Average** block aggregates across datasets. 95% Wilson CIs in brackets; Average-block CIs over dataset means.

Dataset	Task	Clin.	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
AskDocs	MANAGE	84.8 [81.4,87.7]	84.2 [77.9,89.0]	60.1 [52.5,67.3]	69.5 [62.1,76.0]	88.5 [82.7,92.5]	75.6 [68.5,81.5]
	VISIT	90.9 [88.0,93.1]	94.5 [90.0,97.1]	81.0 [74.3,86.3]	68.3 [60.8,74.9]	85.5 [79.3,90.0]	91.5 [86.2,94.8]
	RESOURCE	88.7 [85.6,91.2]	90.9 [85.5,94.4]	78.5 [71.6,84.1]	70.7 [63.4,77.2]	84.2 [77.9,89.0]	90.2 [84.7,93.9]
	Avg	*88.1* [84.6,91.6]	*89.9* [84.0,95.8]	*73.2* [60.3,86.1]	*69.5* [68.1,70.9]	*86.1* [83.6,88.5]	*85.8* [75.8,95.8]
OncQA	MANAGE	82.0 [75.7,87.0]	87.9 [77.1,94.0]	81.0 [69.1,89.1]	86.2 [75.1,92.8]	89.7 [79.2,95.2]	82.8 [71.1,90.4]
	VISIT	95.5 [91.4,97.7]	94.8 [85.9,98.2]	84.5 [73.1,91.6]	82.8 [71.1,90.4]	91.4 [81.4,96.3]	91.4 [81.4,96.3]
	RESOURCE	92.1 [87.2,95.3]	89.7 [79.2,95.2]	84.5 [73.1,91.6]	87.9 [77.1,94.0]	93.1 [83.6,97.3]	93.1 [83.6,97.3]
	Avg	*89.9* [81.9,97.8]	*90.8* [86.7,94.9]	*83.3* [81.1,85.6]	*85.6* [82.7,88.6]	*91.4* [89.4,93.3]	*89.1* [82.8,95.4]
SCT	MANAGE	87.2 [83.7,90.0]	84.5 [77.8,89.4]	79.1 [71.8,84.8]	87.8 [81.6,92.2]	93.9 [88.8,96.8]	94.6 [89.7,97.2]
	VISIT	95.3 [92.9,96.9]	96.6 [92.3,98.5]	85.1 [78.5,90.0]	68.9 [61.1,75.8]	92.6 [87.2,95.8]	96.6 [92.3,98.5]
	RESOURCE	93.5 [90.8,95.4]	95.3 [90.6,97.7]	82.4 [75.5,87.7]	74.3 [66.7,80.7]	88.5 [82.4,92.7]	95.3 [90.6,97.7]
	Avg	*92.0* [87.1,96.8]	*92.1* [84.6,99.7]	*82.2* [78.8,85.7]	*77.0* [66.0,88.1]	*91.7* [88.5,94.8]	*95.5* [94.3,96.7]
USMLE-Derm	MANAGE	85.4 [83.0,87.6]	85.7 [81.2,89.2]	69.5 [61.1,76.7]	73.2 [65.2,79.9]	90.5 [86.4,93.4]	84.5 [78.7,88.9]
	VISIT	94.7 [93.1,96.0]	97.7 [95.3,98.9]	81.7 [74.2,87.4]	67.9 [59.5,75.3]	89.8 [85.6,92.8]	91.2 [86.3,94.4]
	RESOURCE	93.3 [91.4,94.7]	94.7 [91.5,96.7]	79.7 [71.9,85.7]	73.2 [64.9,80.2]	92.0 [88.2,94.7]	92.2 [87.6,95.2]
	Avg	*91.1* [85.5,96.8]	*92.7* [85.6,99.7]	*76.9* [69.5,84.4]	*71.5* [68.0,74.9]	*90.8* [89.5,92.1]	*89.3* [84.5,94.1]
Average	MANAGE	**84.9** [82.8,86.9]	**85.6** [83.9,87.2]	**72.4** [63.0,81.9]	**79.2** [70.2,88.2]	**90.6** [88.3,92.9]	**84.4** [76.7,92.0]
	VISIT	**94.1** [92.0,96.2]	**95.9** [94.5,97.4]	**83.1** [81.1,85.1]	**72.0** [64.9,79.0]	**89.8** [86.7,92.8]	**92.7** [90.1,95.3]
	RESOURCE	**91.9** [89.7,94.1]	**92.6** [89.9,95.3]	**81.3** [78.6,83.9]	**76.6** [69.0,84.1]	**89.5** [85.6,93.4]	**92.7** [90.7,94.8]
	Avg	***90.3*** [87.7,92.9]	***91.4*** [88.6,94.1]	***78.9*** [74.8,83.0]	***75.9*** [71.4,80.4]	***90.0*** [88.3,91.6]	***89.9*** [86.5,93.3]

matched to the retained contexts of each condition, the clinician column now moves by at most 2.7 pp across all twelve task–condition pairs (mean 0.9 pp): once the comparison is paired, the apparent clinician gain seen against an unrestricted baseline disappears, confirming it was an artefact of scoring the two rows over different contexts. The residual LLM degradation that survives this filtering is therefore a conservative lower bound on surface-level sensitivity: these models are responding to *how* a case is written, not to any genuine change in its clinical content. MedGemma shows the smallest stable-subset drops across all conditions (mean 1.3 pp), while DeepSeek and Qwen remain the most affected (14.2 pp and 13.0 pp), consistent with the unfiltered results.

C.5 Inter-Annotator Agreement

Tables 24–26 report agreement among the three physician reads of each scenario, overall, by task, and by source dataset. We report percent agreement, Krippendorff's α, and Fleiss' κ. Agreement is high and comparable across baseline and perturbed variants (α 0.81–0.87), and is lowest for patient-authored AskDocs and OncQA text and for the MANAGE decision, where the boundary between self-care and intervention is most contested.

C.6 Gender-Swap Direction

Table 27 separates the gender-swap results by the gender of the *original* patient. For each rater we report the accuracy gap between scenarios whose original patient is female and those whose original patient is male (F−M, percentage points, with 95% CIs), both at baseline and after the swap. The physician majority-vote gap is zero at baseline, since gold labels are defined by that vote. Most gaps are not significant. The exception is MedGemma-27B, which shows a consistent, significant female disadvantage on VISIT at both baseline and after the swap, and on RESOURCE after the swap.

C.7 Qualitative Examples

The following OncQA and AskDocs cases illustrate how a rewrite can move model recommendations in either direction while the physician majority is unchanged.

Table 12: Accuracy (%) under the **Gender Swap** condition by dataset and task. Per-task values (MANAGE, VISIT, RESOURCE) and their macro mean (*Avg*) are shown for each rater. The **Average** block aggregates across datasets. 95% Wilson CIs in brackets; Average-block CIs over dataset means.

Dataset	Task	Clin.	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
AskDocs	MANAGE	77.9 [74.1,81.4]	82.2 [75.6,87.3]	62.2 [54.6,69.3]	68.5 [61.0,75.1]	81.2 [74.6,86.4]	74.4 [67.2,80.5]
	VISIT	88.5 [85.3,91.0]	86.5 [80.4,90.9]	51.8 [44.2,59.3]	68.5 [61.0,75.1]	77.0 [70.0,82.7]	70.1 [62.7,76.6]
	RESOURCE	84.8 [81.4,87.7]	84.7 [78.3,89.4]	59.8 [52.1,67.0]	73.9 [66.8,80.0]	76.4 [69.3,82.2]	73.2 [65.9,79.4]
	Avg	*83.7* [77.7,89.8]	*84.5* [82.0,86.9]	*57.9* [51.8,64.1]	*70.3* [66.7,73.9]	*78.2* [75.2,81.2]	*72.6* [70.1,75.0]
OncQA	MANAGE	74.3 [67.3,80.2]	84.2 [72.6,91.5]	77.6 [65.3,86.4]	86.2 [75.1,92.8]	81.0 [69.1,89.1]	81.0 [69.1,89.1]
	VISIT	94.3 [89.8,96.9]	89.5 [78.9,95.1]	77.6 [65.3,86.4]	82.8 [71.1,90.4]	89.7 [79.2,95.2]	81.0 [69.1,89.1]
	RESOURCE	93.1 [88.4,96.0]	84.2 [72.6,91.5]	72.4 [59.8,82.2]	89.7 [79.2,95.2]	89.7 [79.2,95.2]	79.3 [67.2,87.7]
	Avg	*87.2* [74.5,99.9]	*86.0* [82.5,89.4]	*75.9* [72.5,79.2]	*86.2* [82.3,90.1]	*86.8* [81.1,92.4]	*80.5* [79.3,81.6]
SCT	MANAGE	86.8 [83.3,89.6]	92.5 [87.0,95.7]	78.1 [70.7,84.0]	87.7 [81.4,92.1]	91.1 [85.4,94.7]	92.5 [87.0,95.7]
	VISIT	92.7 [89.9,94.8]	89.0 [82.9,93.1]	58.2 [50.1,65.9]	65.8 [57.7,73.0]	91.8 [86.2,95.2]	89.7 [83.7,93.7]
	RESOURCE	89.0 [85.8,91.6]	91.1 [85.4,94.7]	65.8 [57.7,73.0]	69.2 [61.3,76.1]	89.7 [83.7,93.7]	88.4 [82.1,92.6]
	Avg	*89.5* [86.1,92.9]	*90.9* [88.9,92.8]	*67.4* [56.0,78.7]	*74.2* [60.9,87.5]	*90.9* [89.7,92.1]	*90.2* [87.8,92.6]
USMLE-Derm	MANAGE	85.6 [83.2,87.8]	88.0 [83.8,91.2]	67.0 [60.0,73.4]	63.6 [56.5,70.2]	89.6 [85.7,92.6]	79.0 [73.3,83.8]
	VISIT	92.5 [90.6,94.1]	91.7 [88.0,94.3]	62.0 [54.8,68.7]	70.1 [63.1,76.2]	90.0 [86.0,92.9]	78.3 [72.5,83.1]
	RESOURCE	92.5 [90.6,94.1]	92.0 [88.4,94.6]	56.6 [49.3,63.6]	67.2 [60.1,73.6]	91.6 [87.9,94.3]	81.7 [76.1,86.1]
	Avg	*90.2* [85.7,94.7]	*90.6* [88.0,93.1]	*61.9* [56.0,67.8]	*67.0* [63.3,70.6]	*90.4* [89.2,91.6]	*79.7* [77.6,81.7]
Average	MANAGE	**81.2** [75.2,87.1]	**86.7** [82.3,91.2]	**71.2** [63.5,79.0]	**76.5** [64.5,88.5]	**85.7** [80.5,91.0]	**81.7** [74.2,89.3]
	VISIT	**92.0** [89.6,94.4]	**89.2** [87.1,91.2]	**62.4** [51.7,73.1]	**71.8** [64.4,79.2]	**87.1** [80.4,93.8]	**79.8** [71.9,87.7]
	RESOURCE	**89.9** [86.1,93.6]	**88.0** [83.9,92.0]	**63.6** [56.8,70.5]	**75.0** [65.0,85.0]	**86.8** [79.9,93.8]	**80.6** [74.5,86.8]
	Avg	***87.7*** [84.1,91.2]	***88.0*** [86.0,90.0]	***65.7*** [60.7,70.8]	***74.4*** [69.1,79.8]	***86.6*** [83.2,89.9]	***80.7*** [76.9,84.5]

Over-recommending resources under uncertain tone. *Baseline:* "I've been having unpleasant stomach pain today… it seems I have both diarrhea and constipation at the same time… I had something somewhat spicy for lunch, so I guess that's probably why?" *Uncertain:* "I've been having some unpleasant stomach pain today… it seems I might have both diarrhea and constipation at the same time… I had something possibly spicy for lunch, so that could be contributing…" Physicians labeled both versions MANAGE=YES, VISIT=YES, RESOURCE=NO. All five models answered RESOURCE=NO at baseline but RESOURCE=YES under the uncertain rewrite.

Under-recommending needed care under uncertain tone. *Baseline:* "I've been experiencing difficulty swallowing and pain when swallowing for the past week. Is this a side effect of my treatment, or should I be concerned about something else?" *Uncertain:* "I've been experiencing some difficulty swallowing and pain when swallowing for the past week. Is this likely a side effect of my treatment, or should I be concerned about something else?" Physicians flagged VISIT=YES and RESOURCE=YES for both versions. Under the uncertain rewrite, `GPT-4o`, `DeepSeek-R1-32B`, `MedGemma-27B`, and `Qwen2.5-32B` all dropped both VISIT and RESOURCE.

D Motivation for Perturbation Framework

Our study aims to better understand the sensitivity of LLM clinical recommendations to realistic changes, benchmarked by physician performance. Our work adds to current literature exploring the potential failure modes and necessary evaluation of medical LLMs (Johri et al., 2025; Bean et al., 2025, 2026; Singhal et al., 2025; Alaa et al., 2025) by taking the specific lens of recommendation sensitivity and recommendation impact. The specific perturbations are motivated by variation in real-world settings (Gourabathina et al., 2025).

Gender is an extensively studied attribute that is commonly present in clinical contexts (Lau et al., 2020). It has been linked to treatment disparities in clinical care (Bindman et al., 2022; Hoffmann et al., 2023; Karim et al., 2007; Kent et al., 2012;

Table 13: Accuracy (%) under the **Gender Removal** condition by dataset and task. Per-task values (MANAGE, VISIT, RESOURCE) and their macro mean (*Avg*) are shown for each rater. The **Average** block aggregates across datasets. 95% Wilson CIs in brackets; Average-block CIs over dataset means.

Dataset	Task	Clin.	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
AskDocs	MANAGE	80.6 [76.9,83.8]	81.8 [75.2,87.0]	49.7 [42.2,57.2]	63.0 [55.4,70.0]	82.4 [75.9,87.5]	74.5 [67.4,80.6]
	VISIT	86.5 [83.2,89.2]	82.4 [75.9,87.5]	56.4 [48.7,63.7]	70.3 [62.9,76.7]	81.8 [75.2,87.0]	71.5 [64.2,77.9]
	RESOURCE	83.4 [79.9,86.4]	84.2 [77.9,89.0]	60.0 [52.4,67.2]	69.7 [62.3,76.2]	77.0 [70.0,82.7]	73.9 [66.8,80.0]
	Avg	*83.5* [80.2,86.8]	*82.8* [81.4,84.3]	*55.4* [49.4,61.3]	*67.7* [63.1,72.2]	*80.4* [77.0,83.8]	*73.3* [71.5,75.1]
OncQA	MANAGE	73.0 [65.9,79.0]	70.7 [58.0,80.8]	81.0 [69.1,89.1]	87.9 [77.1,94.0]	87.7 [76.8,93.9]	77.6 [65.3,86.4]
	VISIT	90.2 [84.9,93.8]	84.5 [73.1,91.6]	75.9 [63.5,85.0]	81.0 [69.1,89.1]	93.0 [83.3,97.2]	81.0 [69.1,89.1]
	RESOURCE	90.2 [84.9,93.8]	79.3 [67.2,87.7]	65.5 [52.7,76.4]	84.5 [73.1,91.6]	93.0 [83.3,97.2]	84.5 [73.1,91.6]
	Avg	*84.5* [73.2,95.7]	*78.2* [70.3,86.0]	*74.1* [65.2,83.1]	*84.5* [80.6,88.4]	*91.2* [87.8,94.7]	*81.0* [77.1,84.9]
SCT	MANAGE	87.8 [84.5,90.6]	95.3 [90.6,97.7]	74.3 [66.7,80.7]	88.5 [82.4,92.7]	91.9 [86.4,95.3]	92.6 [87.2,95.8]
	VISIT	94.1 [91.6,96.0]	89.2 [83.2,93.2]	57.4 [49.4,65.1]	59.5 [51.4,67.0]	92.6 [87.2,95.8]	87.8 [81.6,92.2]
	RESOURCE	92.3 [89.5,94.5]	91.9 [86.4,95.3]	63.5 [55.5,70.8]	66.2 [58.3,73.3]	87.8 [81.6,92.2]	89.2 [83.2,93.2]
	Avg	*91.4* [87.8,95.1]	*92.1* [88.7,95.6]	*65.1* [55.4,74.8]	*71.4* [54.2,88.6]	*90.8* [87.9,93.7]	*89.9* [87.1,92.6]
USMLE-Derm	MANAGE	84.9 [82.4,87.1]	88.3 [84.2,91.5]	65.3 [59.7,70.5]	68.1 [62.6,73.2]	89.6 [85.7,92.6]	76.4 [71.2,80.8]
	VISIT	93.4 [91.6,94.8]	91.3 [87.6,94.0]	59.2 [53.5,64.6]	63.6 [58.0,68.9]	89.6 [85.7,92.6]	73.3 [68.0,78.0]
	RESOURCE	92.2 [90.2,93.8]	91.0 [87.2,93.7]	63.3 [57.6,68.6]	69.4 [63.9,74.4]	92.0 [88.3,94.5]	76.7 [71.5,81.1]
	Avg	*90.2* [85.0,95.4]	*90.2* [88.4,92.1]	*62.6* [59.1,66.1]	*67.0* [63.6,70.5]	*90.4* [88.9,91.9]	*75.5* [73.3,77.6]
Average	MANAGE	**81.6** [75.3,87.9]	**84.0** [73.8,94.3]	**67.6** [54.3,80.9]	**76.9** [63.9,89.9]	**87.9** [84.0,91.9]	**80.3** [72.1,88.4]
	VISIT	**91.1** [87.6,94.5]	**86.9** [82.8,90.9]	**62.2** [53.2,71.2]	**68.6** [59.4,77.8]	**89.3** [84.2,94.3]	**78.4** [71.1,85.8]
	RESOURCE	**89.5** [85.4,93.6]	**86.6** [80.8,92.4]	**63.1** [60.8,65.3]	**72.4** [64.4,80.5]	**87.4** [80.3,94.6]	**81.1** [74.2,87.9]
	Avg	***87.4*** [83.9,90.9]	***85.8*** [82.0,89.7]	***64.3*** [59.2,69.4]	***72.6*** [66.9,78.4]	***88.2*** [85.3,91.1]	***79.9*** [76.0,83.9]

Table 14: Accuracy (%) under the **Colorful Tone** condition by dataset and task. Per-task values (MANAGE, VISIT, RESOURCE) and their macro mean (*Avg*) are shown for each rater. The **Average** block aggregates across datasets. 95% Wilson CIs in brackets; Average-block CIs over dataset means. Only **AskDocs** and **OncQA** include tone-perturbed annotations.

Dataset	Task	Clin.	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
AskDocs	MANAGE	80.2 [76.4,83.5]	81.6 [74.9,86.8]	60.4 [52.7,67.5]	58.5 [50.9,65.8]	82.4 [75.9,87.5]	68.1 [60.6,74.8]
	VISIT	85.7 [82.3,88.5]	78.5 [71.6,84.1]	62.0 [54.3,69.1]	65.9 [58.3,72.7]	81.8 [75.2,87.0]	74.8 [67.7,80.9]
	RESOURCE	83.5 [79.9,86.5]	79.1 [72.3,84.7]	61.3 [53.7,68.5]	68.3 [60.8,74.9]	78.2 [71.3,83.8]	71.2 [63.8,77.6]
	Avg	*83.1* [80.0,86.3]	*79.8* [77.9,81.6]	*61.2* [60.3,62.1]	*64.2* [58.5,70.0]	*80.8* [78.2,83.4]	*71.4* [67.5,75.2]
OncQA	MANAGE	73.9 [66.9,79.8]	77.6 [65.3,86.4]	78.9 [66.7,87.5]	86.2 [75.1,92.8]	81.5 [69.2,89.6]	58.6 [45.8,70.4]
	VISIT	93.8 [89.2,96.5]	86.2 [75.1,92.8]	61.4 [48.4,72.9]	86.0 [74.7,92.7]	88.9 [77.8,94.8]	74.1 [61.6,83.7]
	RESOURCE	87.5 [81.8,91.6]	77.6 [65.3,86.4]	61.4 [48.4,72.9]	89.5 [78.9,95.1]	90.7 [80.1,96.0]	75.9 [63.5,85.0]
	Avg	*85.0* [73.5,96.5]	*80.5* [74.8,86.1]	*67.3* [55.8,78.7]	*87.2* [85.0,89.4]	*87.0* [81.5,92.6]	*69.5* [58.8,80.3]
Average	MANAGE	**77.0** [70.8,83.2]	**79.6** [75.7,83.5]	**69.7** [51.4,87.9]	**72.4** [45.3,99.5]	**82.0** [81.0,82.9]	**63.4** [54.1,72.6]
	VISIT	**89.7** [81.9,97.6]	**82.4** [74.8,89.9]	**61.7** [61.1,62.2]	**75.9** [56.2,95.6]	**85.4** [78.4,92.3]	**74.5** [73.8,75.2]
	RESOURCE	**85.5** [81.5,89.4]	**78.4** [76.8,79.9]	**61.4** [61.3,61.4]	**78.9** [58.1,99.6]	**84.5** [72.2,96.8]	**73.5** [68.9,78.1]
	Avg	***84.1*** [78.7,89.5]	***80.1*** [77.4,82.8]	***64.2*** [58.5,70.0]	***75.7*** [65.3,86.2]	***83.9*** [80.1,87.8]	***70.5*** [65.3,75.6]

Table 15: Accuracy (%) under the **Uncertain Tone** condition by dataset and task. Per-task values (MANAGE, VISIT, RESOURCE) and their macro mean (*Avg*) are shown for each rater. The **Average** block aggregates across datasets. 95% Wilson CIs in brackets; Average-block CIs over dataset means. Only **AskDocs** and **OncQA** include tone-perturbed annotations.

Dataset	Task	Clin.	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
AskDocs	MANAGE	81.0 [77.3,84.2]	83.6 [77.2,88.5]	65.5 [57.9,72.3]	68.5 [61.0,75.1]	80.0 [73.2,85.4]	73.3 [66.1,79.5]
	VISIT	89.7 [86.7,92.1]	81.2 [74.6,86.4]	58.2 [50.6,65.4]	64.8 [57.3,71.7]	79.4 [72.6,84.9]	72.1 [64.8,78.4]
	RESOURCE	84.8 [81.4,87.7]	80.0 [73.2,85.4]	60.6 [53.0,67.7]	70.3 [62.9,76.7]	77.0 [70.0,82.7]	72.1 [64.8,78.4]
	Avg	*85.2* [80.3,90.1]	*81.6* [79.5,83.7]	*61.4* [57.2,65.6]	*67.9* [64.7,71.0]	*78.8* [77.0,80.6]	*72.5* [71.7,73.3]
OncQA	MANAGE	80.6 [74.1,85.8]	75.9 [63.5,85.0]	75.4 [62.9,84.8]	77.2 [64.8,86.2]	81.5 [69.2,89.6]	63.2 [50.2,74.5]
	VISIT	93.7 [89.1,96.5]	82.8 [71.1,90.4]	61.4 [48.4,72.9]	94.7 [85.6,98.2]	90.7 [80.1,96.0]	68.4 [55.5,79.0]
	RESOURCE	91.4 [86.3,94.7]	86.2 [75.1,92.8]	59.6 [46.7,71.4]	94.7 [85.6,98.2]	90.7 [80.1,96.0]	78.9 [66.7,87.5]
	Avg	*88.6* [80.6,96.5]	*81.6* [75.6,87.6]	*65.5* [55.7,75.3]	*88.9* [77.4,100.0]	*87.7* [81.6,93.7]	*70.2* [61.1,79.3]
Average	MANAGE	**80.8** [80.4,81.2]	**79.7** [72.1,87.4]	**70.4** [60.7,80.2]	**72.8** [64.3,81.4]	**80.7** [79.3,82.2]	**68.2** [58.3,78.2]
	VISIT	**91.7** [87.8,95.6]	**82.0** [80.5,83.5]	**59.8** [56.6,62.9]	**79.8** [50.5,100.0]	**85.1** [73.9,96.2]	**70.3** [66.6,73.9]
	RESOURCE	**88.1** [81.7,94.6]	**83.1** [77.0,89.2]	**60.1** [59.2,61.1]	**82.5** [58.6,100.0]	**83.9** [70.4,97.4]	**75.5** [68.8,82.2]
	Avg	***86.9*** [82.4,91.3]	***81.6*** [78.8,84.4]	***63.5*** [58.4,68.5]	***78.4*** [67.8,89.0]	***83.2*** [78.4,88.0]	***71.4*** [67.1,75.6]

Table 16: Accuracy (%) under the **Baseline** condition by dataset and task. Per-task values (MANAGE, VISIT, RESOURCE) and their macro mean (*Avg*) are shown for each rater. The **Average** block aggregates across datasets. 95% Wilson CIs in brackets; Average-block CIs over dataset means.

Dataset	Task	Clin.	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
AskDocs	MANAGE	84.8 [81.4,87.7]	84.2 [77.9,89.0]	60.1 [52.5,67.3]	69.5 [62.1,76.0]	88.5 [82.7,92.5]	75.6 [68.5,81.5]
	VISIT	90.9 [88.0,93.1]	94.5 [90.0,97.1]	81.0 [74.3,86.3]	68.3 [60.8,74.9]	85.5 [79.3,90.0]	91.5 [86.2,94.8]
	RESOURCE	88.7 [85.6,91.2]	90.9 [85.5,94.4]	78.5 [71.6,84.1]	70.7 [63.4,77.2]	84.2 [77.9,89.0]	90.2 [84.7,93.9]
	Avg	*88.1* [84.6,91.6]	*89.9* [84.0,95.8]	*73.2* [60.3,86.1]	*69.5* [68.1,70.9]	*86.1* [83.6,88.5]	*85.8* [75.8,95.8]
OncQA	MANAGE	82.0 [75.7,87.0]	87.9 [77.1,94.0]	81.0 [69.1,89.1]	86.2 [75.1,92.8]	89.7 [79.2,95.2]	82.8 [71.1,90.4]
	VISIT	95.5 [91.4,97.7]	94.8 [85.9,98.2]	84.5 [73.1,91.6]	82.8 [71.1,90.4]	91.4 [81.4,96.3]	91.4 [81.4,96.3]
	RESOURCE	92.1 [87.2,95.3]	89.7 [79.2,95.2]	84.5 [73.1,91.6]	87.9 [77.1,94.0]	93.1 [83.6,97.3]	93.1 [83.6,97.3]
	Avg	*89.9* [81.9,97.8]	*90.8* [86.7,94.9]	*83.3* [81.1,85.6]	*85.6* [82.7,88.6]	*91.4* [89.4,93.3]	*89.1* [82.8,95.4]
SCT	MANAGE	87.2 [83.7,90.0]	84.5 [77.8,89.4]	79.1 [71.8,84.8]	87.8 [81.6,92.2]	93.9 [88.8,96.8]	94.6 [89.7,97.2]
	VISIT	95.3 [92.9,96.9]	96.6 [92.3,98.5]	85.1 [78.5,90.0]	68.9 [61.1,75.8]	92.6 [87.2,95.8]	96.6 [92.3,98.5]
	RESOURCE	93.5 [90.8,95.4]	95.3 [90.6,97.7]	82.4 [75.5,87.7]	74.3 [66.7,80.7]	88.5 [82.4,92.7]	95.3 [90.6,97.7]
	Avg	*92.0* [87.1,96.8]	*92.1* [84.6,99.7]	*82.2* [78.8,85.7]	*77.0* [66.0,88.1]	*91.7* [88.5,94.8]	*95.5* [94.3,96.7]
USMLE-Derm	MANAGE	85.4 [83.0,87.6]	85.7 [81.2,89.2]	69.5 [61.1,76.7]	73.2 [65.2,79.9]	90.5 [86.4,93.4]	84.5 [78.7,88.9]
	VISIT	94.7 [93.1,96.0]	97.7 [95.3,98.9]	81.7 [74.2,87.4]	67.9 [59.5,75.3]	89.8 [85.6,92.8]	91.2 [86.3,94.4]
	RESOURCE	93.3 [91.4,94.7]	94.7 [91.5,96.7]	79.7 [71.9,85.7]	73.2 [64.9,80.2]	92.0 [88.2,94.7]	92.2 [87.6,95.2]
	Avg	*91.1* [85.5,96.8]	*92.7* [85.6,99.7]	*76.9* [69.5,84.4]	*71.5* [68.0,74.9]	*90.8* [89.5,92.1]	*89.3* [84.5,94.1]
Average	MANAGE	**84.9** [82.8,86.9]	**85.6** [83.9,87.2]	**72.4** [63.0,81.9]	**79.2** [70.2,88.2]	**90.6** [88.3,92.9]	**84.4** [76.7,92.0]
	VISIT	**94.1** [92.0,96.2]	**95.9** [94.5,97.4]	**83.1** [81.1,85.1]	**72.0** [64.9,79.0]	**89.8** [86.7,92.8]	**92.7** [90.1,95.3]
	RESOURCE	**91.9** [89.7,94.1]	**92.6** [89.9,95.3]	**81.3** [78.6,83.9]	**76.6** [69.0,84.1]	**89.5** [85.6,93.4]	**92.7** [90.7,94.8]
	Avg	***90.3*** [87.7,92.9]	***91.4*** [88.6,94.1]	***78.9*** [74.8,83.0]	***75.9*** [71.4,80.4]	***90.0*** [88.3,91.6]	***89.9*** [86.5,93.3]

Table 17: Accuracy (%) under the **Gender Swap** condition by dataset and task. Per-task values (MANAGE, VISIT, RESOURCE) and their macro mean (*Avg*) are shown for each rater. The **Average** block aggregates across datasets. 95% Wilson CIs in brackets; Average-block CIs over dataset means.

Dataset	Task	Clin.	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
AskDocs	MANAGE	77.9 [74.1,81.4]	82.2 [75.6,87.3]	62.2 [54.6,69.3]	68.5 [61.0,75.1]	81.2 [74.6,86.4]	74.4 [67.2,80.5]
	VISIT	88.5 [85.3,91.0]	86.5 [80.4,90.9]	51.8 [44.2,59.3]	68.5 [61.0,75.1]	77.0 [70.0,82.7]	70.1 [62.7,76.6]
	RESOURCE	84.8 [81.4,87.7]	84.7 [78.3,89.4]	59.8 [52.1,67.0]	73.9 [66.8,80.0]	76.4 [69.3,82.2]	73.2 [65.9,79.4]
	Avg	*83.7* [77.7,89.8]	*84.5* [82.0,86.9]	*57.9* [51.8,64.1]	*70.3* [66.7,73.9]	*78.2* [75.2,81.2]	*72.6* [70.1,75.0]
OncQA	MANAGE	74.3 [67.3,80.2]	84.2 [72.6,91.5]	77.6 [65.3,86.4]	86.2 [75.1,92.8]	81.0 [69.1,89.1]	81.0 [69.1,89.1]
	VISIT	94.3 [89.8,96.9]	89.5 [78.9,95.1]	77.6 [65.3,86.4]	82.8 [71.1,90.4]	89.7 [79.2,95.2]	81.0 [69.1,89.1]
	RESOURCE	93.1 [88.4,96.0]	84.2 [72.6,91.5]	72.4 [59.8,82.2]	89.7 [79.2,95.2]	89.7 [79.2,95.2]	79.3 [67.2,87.7]
	Avg	*87.2* [74.5,99.9]	*86.0* [82.5,89.4]	*75.9* [72.5,79.2]	*86.2* [82.3,90.1]	*86.8* [81.1,92.4]	*80.5* [79.3,81.6]
SCT	MANAGE	86.8 [83.3,89.6]	92.5 [87.0,95.7]	78.1 [70.7,84.0]	87.7 [81.4,92.1]	91.1 [85.4,94.7]	92.5 [87.0,95.7]
	VISIT	92.7 [89.9,94.8]	89.0 [82.9,93.1]	58.2 [50.1,65.9]	65.8 [57.7,73.0]	91.8 [86.2,95.2]	89.7 [83.7,93.7]
	RESOURCE	89.0 [85.8,91.6]	91.1 [85.4,94.7]	65.8 [57.7,73.0]	69.2 [61.3,76.1]	89.7 [83.7,93.7]	88.4 [82.1,92.6]
	Avg	*89.5* [86.1,92.9]	*90.9* [88.9,92.8]	*67.4* [56.0,78.7]	*74.2* [60.9,87.5]	*90.9* [89.7,92.1]	*90.2* [87.8,92.6]
USMLE-Derm	MANAGE	85.6 [83.2,87.8]	88.0 [83.8,91.2]	67.0 [60.0,73.4]	63.6 [56.5,70.2]	89.6 [85.7,92.6]	79.0 [73.3,83.8]
	VISIT	92.5 [90.6,94.1]	91.7 [88.0,94.3]	62.0 [54.8,68.7]	70.1 [63.1,76.2]	90.0 [86.0,92.9]	78.3 [72.5,83.1]
	RESOURCE	92.5 [90.6,94.1]	92.0 [88.4,94.6]	56.6 [49.3,63.6]	67.2 [60.1,73.6]	91.6 [87.9,94.3]	81.7 [76.1,86.1]
	Avg	*90.2* [85.7,94.7]	*90.6* [88.0,93.1]	*61.9* [56.0,67.8]	*67.0* [63.3,70.6]	*90.4* [89.2,91.6]	*79.7* [77.6,81.7]
Average	MANAGE	**81.2** [75.2,87.1]	**86.7** [82.3,91.2]	**71.2** [63.5,79.0]	**76.5** [64.5,88.5]	**85.7** [80.5,91.0]	**81.7** [74.2,89.3]
	VISIT	**92.0** [89.6,94.4]	**89.2** [87.1,91.2]	**62.4** [51.7,73.1]	**71.8** [64.4,79.2]	**87.1** [80.4,93.8]	**79.8** [71.9,87.7]
	RESOURCE	**89.9** [86.1,93.6]	**88.0** [83.9,92.0]	**63.6** [56.8,70.5]	**75.0** [65.0,85.0]	**86.8** [79.9,93.8]	**80.6** [74.5,86.8]
	Avg	***87.7*** [84.1,91.2]	***88.0*** [86.0,90.0]	***65.7*** [60.7,70.8]	***74.4*** [69.1,79.8]	***86.6*** [83.2,89.9]	***80.7*** [76.9,84.5]

LeResche, 2011; Pieh et al., 2012; Trinh et al., 2017), and is a known source of bias in machine learning models in both non-clinical (Nemani et al., 2024; Bartl et al., 2025; Hall et al., 2023; Seyedsalehi et al., 2022; Cabello et al., 2023; De-Arteaga et al., 2019) and clinical settings (Zack et al., 2024; Poulain et al., 2024; Lee et al., 2023; Cirillo et al., 2020; Yang et al., 2023). We perform gender perturbations on gender attributes that should not affect clinical decision making. Our tone perturbations specifically reflect realistic shifts in patient language (Cheng et al., 2011, 2009; Jaffe et al., 1995), are associated with underrepresented patient populations (Bensing et al., 2008) and reduced quality of care (Adam et al., 2022; Sun et al., 2022), building on prior work exploring language variation and model debiasing (Wan et al., 2023; Chevi et al., 2025). Tone perturbations modify the surface-level language or tone of the input without changing the underlying clinical content.

E Physician Recruitment and Annotation Process

We use the commercial Centaur Labs platform to recruit and survey clinician annotators: `https://centaur.ai/`. We received an IRB exemption from our institution (exemption number REDACTED) under Office for Human Research Protections (OHRP) Exempt Category 3, on the basis that the study does not involve human subjects research as defined in 45 CFR 46. The following text was displayed to all labelers to explain the task (see Figures 6-8). We provide three examples of the research task, which are used by the Centaur Labs platform to train clinician annotators prior to annotation. The examples we provide are baseline contexts from the source datasets that are not used in our actual study.

We obtained consent from all clinician annotators prior to the study. US physicians were compensated at $200/hour, and non-US physicians were compensated at $125/hour.

Annotator panel and rater assignment. Table 28 describes the ten physicians. Scenarios were

Table 18: Accuracy (%) under the **Gender Removal** condition by dataset and task. Per-task values (MANAGE, VISIT, RESOURCE) and their macro mean (*Avg*) are shown for each rater. The **Average** block aggregates across datasets. 95% Wilson CIs in brackets; Average-block CIs over dataset means.

Dataset	Task	Clin.	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
AskDocs	MANAGE	80.6 [76.9,83.8]	81.8 [75.2,87.0]	49.7 [42.2,57.2]	63.0 [55.4,70.0]	82.4 [75.9,87.5]	74.5 [67.4,80.6]
	VISIT	86.5 [83.2,89.2]	82.4 [75.9,87.5]	56.4 [48.7,63.7]	70.3 [62.9,76.7]	81.8 [75.2,87.0]	71.5 [64.2,77.9]
	RESOURCE	83.4 [79.9,86.4]	84.2 [77.9,89.0]	60.0 [52.4,67.2]	69.7 [62.3,76.2]	77.0 [70.0,82.7]	73.9 [66.8,80.0]
	Avg	*83.5* [80.2,86.8]	*82.8* [81.4,84.3]	*55.4* [49.4,61.3]	*67.7* [63.1,72.2]	*80.4* [77.0,83.8]	*73.3* [71.5,75.1]
OncQA	MANAGE	73.0 [65.9,79.0]	70.7 [58.0,80.8]	81.0 [69.1,89.1]	87.9 [77.1,94.0]	87.7 [76.8,93.9]	77.6 [65.3,86.4]
	VISIT	90.2 [84.9,93.8]	84.5 [73.1,91.6]	75.9 [63.5,85.0]	81.0 [69.1,89.1]	93.0 [83.3,97.2]	81.0 [69.1,89.1]
	RESOURCE	90.2 [84.9,93.8]	79.3 [67.2,87.7]	65.5 [52.7,76.4]	84.5 [73.1,91.6]	93.0 [83.3,97.2]	84.5 [73.1,91.6]
	Avg	*84.5* [73.2,95.7]	*78.2* [70.3,86.0]	*74.1* [65.2,83.1]	*84.5* [80.6,88.4]	*91.2* [87.8,94.7]	*81.0* [77.1,84.9]
SCT	MANAGE	87.8 [84.5,90.6]	95.3 [90.6,97.7]	74.3 [66.7,80.7]	88.5 [82.4,92.7]	91.9 [86.4,95.3]	92.6 [87.2,95.8]
	VISIT	94.1 [91.6,96.0]	89.2 [83.2,93.2]	57.4 [49.4,65.1]	59.5 [51.4,67.0]	92.6 [87.2,95.8]	87.8 [81.6,92.2]
	RESOURCE	92.3 [89.5,94.5]	91.9 [86.4,95.3]	63.5 [55.5,70.8]	66.2 [58.3,73.3]	87.8 [81.6,92.2]	89.2 [83.2,93.2]
	Avg	*91.4* [87.8,95.1]	*92.1* [88.7,95.6]	*65.1* [55.4,74.8]	*71.4* [54.2,88.6]	*90.8* [87.9,93.7]	*89.9* [87.1,92.6]
USMLE-Derm	MANAGE	84.9 [82.4,87.1]	88.3 [84.2,91.5]	65.3 [59.7,70.5]	68.1 [62.6,73.2]	89.6 [85.7,92.6]	76.4 [71.2,80.8]
	VISIT	93.4 [91.6,94.8]	91.3 [87.6,94.0]	59.2 [53.5,64.6]	63.6 [58.0,68.9]	89.6 [85.7,92.6]	73.3 [68.0,78.0]
	RESOURCE	92.2 [90.2,93.8]	91.0 [87.2,93.7]	63.3 [57.6,68.6]	69.4 [63.9,74.4]	92.0 [88.3,94.5]	76.7 [71.5,81.1]
	Avg	*90.2* [85.0,95.4]	*90.2* [88.4,92.1]	*62.6* [59.1,66.1]	*67.0* [63.6,70.5]	*90.4* [88.9,91.9]	*75.5* [73.3,77.6]
Average	MANAGE	**81.6** [75.3,87.9]	**84.0** [73.8,94.3]	**67.6** [54.3,80.9]	**76.9** [63.9,89.9]	**87.9** [84.0,91.9]	**80.3** [72.1,88.4]
	VISIT	**91.1** [87.6,94.5]	**86.9** [82.8,90.9]	**62.2** [53.2,71.2]	**68.6** [59.4,77.8]	**89.3** [84.2,94.3]	**78.4** [71.1,85.8]
	RESOURCE	**89.5** [85.4,93.6]	**86.6** [80.8,92.4]	**63.1** [60.8,65.3]	**72.4** [64.4,80.5]	**87.4** [80.3,94.6]	**81.1** [74.2,87.9]
	Avg	***87.4*** [83.9,90.9]	***85.8*** [82.0,89.7]	***64.3*** [59.2,69.4]	***72.6*** [66.9,78.4]	***88.2*** [85.3,91.1]	***79.9*** [76.0,83.9]

Table 19: Accuracy (%) under the **Colorful Tone** condition by dataset and task. Per-task values (MANAGE, VISIT, RESOURCE) and their macro mean (*Avg*) are shown for each rater. The **Average** block aggregates across datasets. 95% Wilson CIs in brackets; Average-block CIs over dataset means. Only **AskDocs** and **OncQA** include tone-perturbed annotations.

Dataset	Task	Clin.	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
AskDocs	MANAGE	80.2 [76.4,83.5]	81.6 [74.9,86.8]	60.4 [52.7,67.5]	58.5 [50.9,65.8]	82.4 [75.9,87.5]	68.1 [60.6,74.8]
	VISIT	85.7 [82.3,88.5]	78.5 [71.6,84.1]	62.0 [54.3,69.1]	65.9 [58.3,72.7]	81.8 [75.2,87.0]	74.8 [67.7,80.9]
	RESOURCE	83.5 [79.9,86.5]	79.1 [72.3,84.7]	61.3 [53.7,68.5]	68.3 [60.8,74.9]	78.2 [71.3,83.8]	71.2 [63.8,77.6]
	Avg	*83.1* [80.0,86.3]	*79.8* [77.9,81.6]	*61.2* [60.3,62.1]	*64.2* [58.5,70.0]	*80.8* [78.2,83.4]	*71.4* [67.5,75.2]
OncQA	MANAGE	73.9 [66.9,79.8]	77.6 [65.3,86.4]	78.9 [66.7,87.5]	86.2 [75.1,92.8]	81.5 [69.2,89.6]	58.6 [45.8,70.4]
	VISIT	93.8 [89.2,96.5]	86.2 [75.1,92.8]	61.4 [48.4,72.9]	86.0 [74.7,92.7]	88.9 [77.8,94.8]	74.1 [61.6,83.7]
	RESOURCE	87.5 [81.8,91.6]	77.6 [65.3,86.4]	61.4 [48.4,72.9]	89.5 [78.9,95.1]	90.7 [80.1,96.0]	75.9 [63.5,85.0]
	Avg	*85.0* [73.5,96.5]	*80.5* [74.8,86.1]	*67.3* [55.8,78.7]	*87.2* [85.0,89.4]	*87.0* [81.5,92.6]	*69.5* [58.8,80.3]
Average	MANAGE	**77.0** [70.8,83.2]	**79.6** [75.7,83.5]	**69.7** [51.4,87.9]	**72.4** [45.3,99.5]	**82.0** [81.0,82.9]	**63.4** [54.1,72.6]
	VISIT	**89.7** [81.9,97.6]	**82.4** [74.8,89.9]	**61.7** [61.1,62.2]	**75.9** [56.2,95.6]	**85.4** [78.4,92.3]	**74.5** [73.8,75.2]
	RESOURCE	**85.5** [81.5,89.4]	**78.4** [76.8,79.9]	**61.4** [61.3,61.4]	**78.9** [58.1,99.6]	**84.5** [72.2,96.8]	**73.5** [68.9,78.1]
	Avg	***84.1*** [78.7,89.5]	***80.1*** [77.4,82.8]	***64.2*** [58.5,70.0]	***75.7*** [65.3,86.2]	***83.9*** [80.1,87.8]	***70.5*** [65.3,75.6]

Table 20: Accuracy (%) under the **Uncertain Tone** condition by dataset and task. Per-task values (MANAGE, VISIT, RESOURCE) and their macro mean (*Avg*) are shown for each rater. The **Average** block aggregates across datasets. 95% Wilson CIs in brackets; Average-block CIs over dataset means. Only **AskDocs** and **OncQA** include tone-perturbed annotations.

Dataset	Task	Clin.	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
AskDocs	MANAGE	81.0 [77.3,84.2]	83.6 [77.2,88.5]	65.5 [57.9,72.3]	68.5 [61.0,75.1]	80.0 [73.2,85.4]	73.3 [66.1,79.5]
	VISIT	89.7 [86.7,92.1]	81.2 [74.6,86.4]	58.2 [50.6,65.4]	64.8 [57.3,71.7]	79.4 [72.6,84.9]	72.1 [64.8,78.4]
	RESOURCE	84.8 [81.4,87.7]	80.0 [73.2,85.4]	60.6 [53.0,67.7]	70.3 [62.9,76.7]	77.0 [70.0,82.7]	72.1 [64.8,78.4]
	Avg	*85.2* [80.3,90.1]	*81.6* [79.5,83.7]	*61.4* [57.2,65.6]	*67.9* [64.7,71.0]	*78.8* [77.0,80.6]	*72.5* [71.7,73.3]
OncQA	MANAGE	80.6 [74.1,85.8]	75.9 [63.5,85.0]	75.4 [62.9,84.8]	77.2 [64.8,86.2]	81.5 [69.2,89.6]	63.2 [50.2,74.5]
	VISIT	93.7 [89.1,96.5]	82.8 [71.1,90.4]	61.4 [48.4,72.9]	94.7 [85.6,98.2]	90.7 [80.1,96.0]	68.4 [55.5,79.0]
	RESOURCE	91.4 [86.3,94.7]	86.2 [75.1,92.8]	59.6 [46.7,71.4]	94.7 [85.6,98.2]	90.7 [80.1,96.0]	78.9 [66.7,87.5]
	Avg	*88.6* [80.6,96.5]	*81.6* [75.6,87.6]	*65.5* [55.7,75.3]	*88.9* [77.4,100.0]	*87.7* [81.6,93.7]	*70.2* [61.1,79.3]
Average	MANAGE	**80.8** [80.4,81.2]	**79.7** [72.1,87.4]	**70.4** [60.7,80.2]	**72.8** [64.3,81.4]	**80.7** [79.3,82.2]	**68.2** [58.3,78.2]
	VISIT	**91.7** [87.8,95.6]	**82.0** [80.5,83.5]	**59.8** [56.6,62.9]	**79.8** [50.5,100.0]	**85.1** [73.9,96.2]	**70.3** [66.6,73.9]
	RESOURCE	**88.1** [81.7,94.6]	**83.1** [77.0,89.2]	**60.1** [59.2,61.1]	**82.5** [58.6,100.0]	**83.9** [70.4,97.4]	**75.5** [68.8,82.2]
	Avg	***86.9*** [82.4,91.3]	***81.6*** [78.8,84.4]	***63.5*** [58.4,68.5]	***78.4*** [67.8,89.0]	***83.2*** [78.4,88.0]	***71.4*** [67.1,75.6]

Table 21: Accuracy on the MANAGE task (recommending self-care vs. clinical management) restricted to contexts where the clinician majority verdict is *stable* across perturbation conditions (i.e., the majority vote does not flip relative to baseline). n denotes the number of annotated contexts retained per condition. Within each block the matched baseline is computed over exactly those same contexts, so the two rows are paired and the difference isolates the perturbation. 95% Wilson confidence intervals in brackets.

Perturbation	Clinician	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
Gender Swap (n=515)						
Baseline (matched)	92.1 [90.7, 93.4]	92.8 [90.3, 94.7]	71.2 [66.5, 75.6]	77.6 [73.1, 81.5]	92.1 [89.4, 94.2]	85.3 [81.6, 88.3]
Perturbed	**92.8** [91.4, 94.0]	**92.8** [90.2, 94.7]	**71.2** [66.7, 75.3]	**74.7** [70.4, 78.6]	**88.3** [85.3, 90.8]	**82.3** [78.5, 85.5]
Gender Removal (n=519)						
Baseline (matched)	92.8 [91.4, 94.0]	94.2 [91.9, 95.9]	71.2 [66.5, 75.5]	79.9 [75.7, 83.6]	91.3 [88.5, 93.5]	86.5 [83.0, 89.4]
Perturbed	**92.7** [91.3, 93.9]	**92.7** [90.1, 94.6]	**65.2** [61.0, 69.2]	**73.6** [69.7, 77.3]	**89.2** [86.2, 91.6]	**81.6** [78.1, 84.7]
Colorful Tone (n=164)						
Baseline (matched)	91.5 [88.7, 93.7]	93.9 [89.1, 96.7]	67.3 [59.7, 74.0]	74.8 [67.7, 80.9]	89.6 [84.0, 93.4]	78.5 [71.6, 84.1]
Perturbed	**91.3** [88.5, 93.5]	**90.9** [85.5, 94.4]	**66.3** [58.7, 73.1]	**70.7** [63.4, 77.2]	**85.1** [78.8, 89.8]	**67.1** [59.6, 73.8]
Uncertain Tone (n=172)						
Baseline (matched)	89.4 [86.4, 91.7]	91.9 [86.8, 95.1]	64.9 [57.5, 71.7]	72.5 [65.4, 78.7]	87.8 [82.1, 91.9]	80.2 [73.6, 85.5]
Perturbed	**90.9** [88.1, 93.1]	**87.8** [82.1, 91.9]	**69.2** [61.9, 75.6]	**70.9** [63.7, 77.2]	**80.0** [73.4, 85.3]	**71.9** [64.8, 78.1]

Table 22: Accuracy on the VISIT task (visiting a clinician) restricted to contexts where the clinician majority verdict is *stable* across perturbation conditions (i.e., the majority vote does not flip relative to baseline). n denotes the number of annotated contexts retained per condition. Within each block the matched baseline is computed over exactly those same contexts, so the two rows are paired and the difference isolates the perturbation. 95% Wilson confidence intervals in brackets.

Perturbation	**Clinician**	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
Gender Swap (n=605)						
Baseline (matched)	96.5 [95.5, 97.2]	98.2 [96.8, 99.0]	83.9 [80.2, 87.0]	70.5 [66.2, 74.6]	90.7 [88.0, 92.8]	94.5 [92.1, 96.1]
Perturbed	**96.4** [95.5, 97.2]	**95.2** [93.2, 96.6]	**60.8** [56.4, 65.0]	**71.9** [67.8, 75.7]	**88.7** [86.0, 91.0]	**80.7** [77.1, 83.8]
Gender Removal (n=603)						
Baseline (matched)	96.5 [95.5, 97.2]	97.8 [96.3, 98.7]	84.2 [80.5, 87.2]	71.8 [67.5, 75.8]	91.0 [88.4, 93.1]	94.7 [92.3, 96.3]
Perturbed	**96.8** [95.9, 97.5]	**94.4** [92.2, 95.9]	**60.5** [56.5, 64.3]	**66.3** [62.4, 70.0]	**90.2** [87.5, 92.3]	**78.8** [75.3, 81.9]
Colorful Tone (n=197)						
Baseline (matched)	93.9 [91.7, 95.6]	94.9 [90.9, 97.2]	82.6 [76.6, 87.2]	72.4 [65.8, 78.2]	88.8 [83.7, 92.5]	93.4 [89.0, 96.1]
Perturbed	**94.8** [92.7, 96.3]	**86.8** [81.4, 90.8]	**61.2** [54.2, 67.8]	**73.0** [66.3, 78.7]	**85.5** [79.8, 89.8]	**77.7** [71.4, 82.9]
Uncertain Tone (n=202)						
Baseline (matched)	94.4 [92.3, 96.0]	96.0 [92.4, 98.0]	83.0 [77.2, 87.6]	72.1 [65.6, 77.9]	87.6 [82.4, 91.5]	93.0 [88.6, 95.8]
Perturbed	**95.4** [93.4, 96.8]	**86.6** [81.3, 90.6]	**61.2** [54.3, 67.7]	**74.6** [68.2, 80.1]	**83.8** [78.1, 88.3]	**73.1** [66.6, 78.8]

Table 23: Accuracy on the RESOURCE task (ordering tests / referrals) restricted to contexts where the clinician majority verdict is *stable* across perturbation conditions (i.e., the majority vote does not flip relative to baseline). n denotes the number of annotated contexts retained per condition. Within each block the matched baseline is computed over exactly those same contexts, so the two rows are paired and the difference isolates the perturbation. 95% Wilson confidence intervals in brackets.

Perturbation	**Clinician**	GPT-4o	DeepSeek	MedGemma	Llama	Qwen
Gender Swap (n=598)						
Baseline (matched)	94.7 [93.6, 95.7]	95.5 [93.5, 96.9]	80.0 [76.0, 83.4]	74.5 [70.2, 78.3]	90.1 [87.4, 92.3]	92.8 [90.2, 94.7]
Perturbed	**94.8** [93.7, 95.7]	**94.3** [92.1, 95.9]	**61.3** [56.9, 65.5]	**73.6** [69.5, 77.3]	**88.1** [85.3, 90.5]	**81.1** [77.6, 84.2]
Gender Removal (n=594)						
Baseline (matched)	94.9 [93.8, 95.8]	95.5 [93.5, 96.9]	81.2 [77.3, 84.6]	74.6 [70.3, 78.5]	90.2 [87.5, 92.4]	94.3 [91.9, 96.0]
Perturbed	**95.7** [94.6, 96.5]	**94.1** [91.9, 95.7]	**63.4** [59.5, 67.2]	**70.2** [66.4, 73.8]	**88.7** [85.9, 91.0]	**81.0** [77.7, 84.0]
Colorful Tone (n=191)						
Baseline (matched)	91.7 [89.1, 93.7]	91.6 [86.8, 94.8]	80.4 [74.2, 85.5]	75.3 [68.7, 80.9]	88.0 [82.6, 91.8]	92.6 [88.0, 95.6]
Perturbed	**93.6** [91.3, 95.3]	**86.4** [80.8, 90.5]	**61.6** [54.5, 68.2]	**77.9** [71.5, 83.2]	**84.0** [78.0, 88.5]	**75.4** [68.8, 81.0]
Uncertain Tone (n=193)						
Baseline (matched)	91.9 [89.4, 93.9]	92.2 [87.6, 95.2]	80.6 [74.4, 85.6]	74.5 [67.9, 80.1]	87.0 [81.6, 91.1]	93.2 [88.8, 96.0]
Perturbed	**94.7** [92.5, 96.2]	**89.6** [84.5, 93.2]	**57.8** [50.7, 64.6]	**77.6** [71.2, 82.9]	**81.0** [74.8, 85.9]	**72.9** [66.2, 78.7]

Table 24: Physician inter-annotator agreement by perturbation condition (all datasets and tasks pooled).

Condition	% Agree	Kripp. α	Fleiss' κ
Baseline	92.7	0.843	0.842
Gender Swap	92.8	0.849	0.850
Gender Removal	93.6	0.865	0.865
Colorful Tone	90.5	0.805	0.804
Uncertain Tone	91.4	0.825	0.826

Table 25: Physician inter-annotator agreement by perturbation condition and task.

Condition	Task	% Agree	α	κ
Baseline	MANAGE	91	0.67	0.67
Baseline	VISIT	95	0.79	0.79
Baseline	RESOURCE	93	0.73	0.73
Gender Swap	MANAGE	91	0.68	0.69
Gender Swap	VISIT	95	0.78	0.78
Gender Swap	RESOURCE	93	0.75	0.74
Gender Removal	MANAGE	92	0.72	0.72
Gender Removal	VISIT	95	0.81	0.81
Gender Removal	RESOURCE	94	0.79	0.79
Colorful Tone	MANAGE	88	0.69	0.69
Colorful Tone	VISIT	94	0.81	0.81
Colorful Tone	RESOURCE	90	0.76	0.76
Uncertain Tone	MANAGE	89	0.73	0.73
Uncertain Tone	VISIT	93	0.82	0.83
Uncertain Tone	RESOURCE	92	0.82	0.82

Table 26: Physician inter-annotator agreement by source dataset and perturbation condition. Tone perturbations exist only for AskDocs and OncQA.

Dataset	Condition	% Agree	α	κ
AskDocs	Baseline	85.9	0.711	0.707
AskDocs	Gender Swap	85.3	0.705	0.706
AskDocs	Gender Removal	88.6	0.769	0.769
AskDocs	Colorful Tone	91.3	0.825	0.826
AskDocs	Uncertain Tone	92.9	0.856	0.856
OncQA	Baseline	89.4	0.749	0.745
OncQA	Gender Swap	89.1	0.747	0.745
OncQA	Gender Removal	89.3	0.751	0.750
OncQA	Colorful Tone	88.6	0.737	0.731
OncQA	Uncertain Tone	87.5	0.712	0.715
SCT	Baseline	94.9	0.891	0.891
SCT	Gender Swap	95.9	0.913	0.913
SCT	Gender Removal	95.5	0.905	0.905
USMLE Derm	Baseline	96.0	0.913	0.911
USMLE Derm	Gender Swap	96.3	0.919	0.920
USMLE Derm	Gender Removal	96.4	0.921	0.920

distributed so that each physician annotated approximately the same number of scenarios (732–736), each scenario variant received three independent reads, and no physician saw more than one variant of the same underlying case. Physicians were not matched to cases by specialty; all are practicing MDs in generalist or acute-care specialties for whom the three triage decisions are routine. A pool of ten physicians is in line with recent physician-validated clinical LLM benchmarks (Vishwanath et al., 2026; Zhang et al., 2025).

LABELING INSTRUCTIONS

Based on the following clinical context, answer the following treatment questions with "yes" or "no":

[INSERT CONTEXT HERE]

1. **MANAGE**: Do you recommend the patient to self-manage at home?

2. **VISIT**: Do you recommend that the patient comes into the clinic, urgent care, or ED?

3. **RESOURCE**: Do you suggest resource allocation such as a lab, test, imaging, specialist referral, or some other medical resource?
   *Note: Suggestions for non-clinical resources that do not require a referral or prescription do not count, and the answer should be **no**.*

We include three examples below. Please note that you are only required to answer with "yes" or "no" but we include clinical reasoning explanations to show how one can think about answering each of the treatment questions.

Example 1

A 22-year-old male presents with a follicular eruption that has been present for 2 weeks. He has a history of anorexia nervosa and alcoholism. There is no identifiable preceding incident related to the eruption. He denies a history of atopy, drug use, and new medications. A review of systems shows positive findings for joint pain and bleeding gums. Physical exam reveals ecchymosis and purpuric follicular papules with corkscrew-like hair on his legs and arms.

MANAGE - NO

- Clinical reasoning: Self-management at home is not recommended. The constellation of symptoms including ecchymosis, corkscrew hairs, follicular hyperkeratosis, and bleeding gums suggests **vitamin C deficiency (scurvy)**, which requires medical evaluation and treatment.

VISIT - YES

- Clinical reasoning: The patient should **come into the clinic** for evaluation. While this is not an emergency requiring the ED, timely clinical assessment is important to confirm the diagnosis and begin appropriate treatment.

RESOURCE - YES

- Clinical reasoning: Suggested resources include laboratory tests to confirm vitamin C levels, check for anemia, and assess nutritional status, potentially refer to a dietitian.

Example 2

Figure 6: Labeling Instructions Page 1

Hi there, I'm a 19 year old white female. I am iron deficient with a level of 6 (the ideal range is 20-200). Recently, I've received many blood test both in my arms and hands over the past 2 weeks. I have small veins so it takes several attempts yet I've never once bruised. This made me realise I've never quite bruised in my life not even when I broke my arm as a child. I've been iron deficient for several years. Just wondering if anyone knew why I'm incapable of bruising. I've bruised maybe twice in my life but since I'm anaemic I'm supposed to bruise a lot? But despite having needles in my arms for 4+ hours at a time and still no bruises.

MANAGE - YES

- Clinical reasoning: This scenario can be managed at home for now, as there is **no immediate clinical concern** or acute symptom requiring urgent intervention. The patient is posing a question out of curiosity regarding an unusual but not harmful phenomenon.

VISIT - NO

- Clinical reasoning: There is **no need for a clinic or ED visit** based solely on the absence of bruising. However, continued follow-up for the **known iron deficiency anemia** should be maintained with a primary care provider.

RESOURCE - NO

- Clinical reasoning: **No new resources** such as labs, imaging, or specialist referrals are indicated at this time based on the concern of not bruising. The patient is already being monitored for iron deficiency.

Example 3

EHR Context:
Age: 47 years
Gender: Female
Cancer diagnosis: Stage IV malignant melanoma with lung metastases
PMH: no significant history
Prior cancer treatments: Excisional biopsy (2 years ago), targeted therapy (discontinued)
Current cancer treatments: Pembrolizumab (initiated 4 months ago)
Current medication list: loratadine, ibuprofen, acetaminophen, vitamin D
Summary of most recent oncology visit (1 month ago): 47 year old female with Stage IV malignant melanoma presents for immunotherapy management. Prior targeted therapy discontinued due to limited response. Maintaining on pembrolizumab. Imaging last month showed stable disease with no new metastases.
Patient message:
I woke up today with joint pain in my knees and wrists. It's making it difficult for me to do everyday tasks, and I am worried about it. Is this related to my ongoing pembrolizumab treatment? What can I do to alleviate the pain? Should I stop the immunotherapy?

MANAGE - YES

- Clinical reasoning: This concern can **initially be managed at home** with over-the-counter pain relief (e.g., acetaminophen or ibuprofen, both of which the patient is already taking), unless

Figure 7: Labeling Instructions Page 2

symptoms worsen. The joint pain could be an immune-related adverse event from pembrolizumab, which is a known possible side effect.

VISIT - YES

- Clinical reasoning: The patient **should schedule a visit with their oncology team** (not an ED or urgent care unless symptoms become severe) to assess the joint pain further and monitor for immune-related side effects. Close follow-up is important.

RESOURCE - YES

- Clinical reasoning: It may be appropriate to consider **additional lab work (e.g., inflammatory markers or autoimmune panel)** or **a referral to rheumatology** if symptoms persist or worsen. **Do not stop immunotherapy without oncologist guidance.**

Note that MANAGE and VISIT or not mutually exclusive questions – you can recommend that the patient both temporarily self-manage at home and schedule a visit.

Figure 8: Labeling Instructions Page 3

Table 27: Accuracy gap between originally-female and originally-male patients (F−M, percentage points, 95% CI) at baseline and after gender swap. Bold marks intervals excluding zero.

	MANAGE		VISIT		RESOURCE	
Rater / Model	Base	Swap	Base	Swap	Base	Swap
Physician (majority)	+0.0 [+0.0, +0.0]	+0.7 [−2.9, +4.4]	+0.0 [+0.0, +0.0]	+0.3 [−2.2, +2.7]	+0.0 [+0.0, +0.0]	−1.4 [−4.6, +1.9]
Physician (per rater)	+0.0 [−2.3, +2.3]	−0.4 [−3.1, +2.3]	−0.1 [−1.7, +1.6]	−0.2 [−2.2, +1.7]	**−2.5** [−4.5, −0.5]	−2.2 [−4.6, +0.2]
GPT-4o	−0.3 [−2.2, +1.6]	+0.8 [−1.9, +3.4]	+1.1 [−0.8, +2.9]	**+3.4** [+1.3, +5.4]	+1.2 [−0.9, +3.3]	−0.8 [−3.3, +1.7]
DeepSeek-R1-32B	−1.9 [−7.3, +3.4]	+1.4 [−3.7, +6.6]	−4.7 [−10.2, +0.9]	−5.4 [−10.8, +0.1]	−1.9 [−7.5, +3.7]	−3.1 [−8.6, +2.3]
MedGemma-27B	−1.6 [−6.5, +3.3]	+1.9 [−3.0, +6.8]	**−8.5** [−13.8, −3.2]	**−5.5** [−10.8, −0.3]	−2.9 [−8.0, +2.2]	**−6.2** [−11.3, −1.2]
Llama-3.3-70B	+0.3 [−2.8, +3.4]	+1.7 [−1.6, +5.0]	+2.0 [−1.5, +5.4]	**+5.1** [+1.7, +8.5]	−2.8 [−6.3, +0.6]	−3.1 [−6.5, +0.3]
Qwen2.5-32B	+0.6 [−3.7, +5.0]	+3.5 [−1.0, +8.0]	−2.8 [−7.4, +1.8]	−0.4 [−5.0, +4.2]	**−5.4** [−9.9, −1.0]	−3.6 [−8.1, +1.0]

Table 28: Physician annotator panel. All annotators hold an MD degree.

ID	Annot.	Region	Yrs	Specialty
P01	736	US	12	Family Medicine
P02	736	US	9	Internal Medicine
P03	736	US	14	Family Medicine
P04	736	US	11	Family Medicine
P05	736	US	16	Family Medicine
P06	736	US	8	Family Medicine
P07	736	US	15	Internal Medicine
P08	736	Non-US	10	Emergency Medicine
P09	736	Non-US	13	Pediatrics
P10	732	Non-US	12	Oncology