Large Language Models for Cancer Communication: Evaluating Linguistic Quality, Safety, and Accessibility in Generative AI

Agnik Saha, Victoria Churchill, Anny D. Rodriguez, Ugur Kursuncu, Muhammed Y. Idris

Submitted to: Journal of Medical Internet Research
on: August 25, 2025

Table of Contents

Large Language Models for Cancer Communication: Evaluating Linguistic Quality, Safety, and Accessibility in Generative AI

Agnik Saha[1]; Victoria Churchill[2]; Anny D. Rodriguez[2]; Ugur Kursuncu[1]; Muhammed Y. Idris[2]

[1] Georgia State University Atlanta US
[2] Morehouse School of Medicine Atlanta US

Corresponding Author:
Muhammed Y. Idris

Morehouse School of Medicine
720 Westview Drive, S.W.
Atlanta
US

Abstract

Background: Effective communication about breast and cervical cancers remains a public health challenge, with widespread misinformation and barriers to cancer-related language understanding. Large Language Models (LLMs) offer potential for scalable health communication, yet tradeoffs between quality, safety, and accessibility of general-purpose and medical-domain LLMs remain underexplored.

Objective: We propose a comprehensive evaluation framework and systematically assesses the performance of LLMs in generating breast and cervical cancer information, with a focus on linguistic quality, safety and trustworthiness, and communication accessibility and affectiveness

Methods: This mixed-methods evaluation study assessed outputs from five general-purpose and three medical large language models (LLMs) using real-world breast and cervical cancer–related questions curated from publicly available medical datasets. LLM-generated responses were evaluated in a controlled offline setting. Primary outcomes included linguistic quality (fluency, coherence, accuracy), safety and trustworthiness (toxicity, bias, harm potential), and communication accessibility and affectiveness (readability, empathy, clarity). Qualitative ratings were performed by domain experts, while quantitative metrics were compared across models. Statistical analyses included Welch's ANOVA to detect differences in metric scores, Games-Howell tests for pairwise comparisons, and Hedges' g to assess effect sizes.

Results: General-purpose LLMs, particularly Llama 3 and Gemma, demonstrated superior linguistic quality and affectiveness but often produced complex outputs that may limit accessibility. In contrast, medical LLMs (e.g., MedAlpaca, BioMistral) generated simpler content suitable for broader audiences but scored lower in safety and empathy due to higher levels of hallucination, bias, and toxicity.

Conclusions: While LLMs show promise for improving digital cancer communication, our findings reveal a trade-off between domain specialization and overall communication quality and safety. Future development of health-focused LLMs should prioritize hybrid modeling strategies to enhance trust, clarity, and clinical relevance in patient-facing tools. Clinical Trial: Not applicable

(JMIR Preprints 25/08/2025:82971)
DOI: https://doi.org/10.2196/preprints.82971

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Title: Large Language Models for Cancer Communication: Evaluating Linguistic Quality, Safety, and Accessibility in Generative AI

Authors: Agnik Saha, Victoria Churchill, Anny D. Rodriguez, Ugur Kursuncu, Muhammed Y. Idris

Corresponding Author: Muhammed Y. Idris, PhD

ABSTRACT

Background:

Effective communication about breast and cervical cancers remains a public health challenge, with widespread misinformation and barriers to cancer-related language understanding. Large Language Models (LLMs) offer potential for scalable health communication, yet tradeoffs between quality, safety, and accessibility of general-purpose and medical-domain LLMs remain underexplored.

Objective:

We propose a comprehensive evaluation framework and systematically assesses the performance of LLMs in generating breast and cervical cancer information, with a focus on linguistic quality, safety and trustworthiness, and communication accessibility and affectiveness.

Methods:

This mixed-methods evaluation study assessed outputs from five general-purpose and three medical large language models (LLMs) using real-world breast and cervical cancer–related questions curated from publicly available medical datasets. LLM-generated responses were evaluated in a controlled offline setting. Primary outcomes included linguistic quality (fluency, coherence, accuracy), safety and trustworthiness (toxicity, bias, harm potential), and communication accessibility and affectiveness (readability, empathy, clarity). Qualitative ratings were performed by domain experts,

while quantitative metrics were compared across models. Statistical analyses included Welch's ANOVA to detect differences in metric scores, Games-Howell tests for pairwise comparisons, and Hedges' g to assess effect sizes.

Results:

General-purpose LLMs, particularly Llama 3 and Gemma, demonstrated superior linguistic quality and affectiveness but often produced complex outputs that may limit accessibility. In contrast, medical LLMs (e.g., MedAlpaca, BioMistral) generated simpler content suitable for broader audiences but scored lower in safety and empathy due to higher levels of hallucination, bias, and toxicity.

Conclusions:

While LLMs show promise for improving digital cancer communication, our findings reveal a trade-off between domain specialization and overall communication quality and safety. Future development of health-focused LLMs should prioritize hybrid modeling strategies to enhance trust, clarity, and clinical relevance in patient-facing tools.

Keywords: Large Language Models; Artificial Intelligence; Natural Language Processing; Medical Informatics; Health Communication; Breast Neoplasms; Uterine Cervical Neoplasms

INTRODUCTION

Cancer remains a leading cause of morbidity and mortality among women in the U.S., making it a critical public health issue. Breast cancer is the most commonly diagnosed cancer among women, with an estimated 310, 720 new cases and 42, 250 deaths projected in 2024.[1] Despite improvements in screening and treatment, disparities in cancer outcomes persist. For instance, Black women experience a 40% higher breast cancer mortality rate than White women, despite similar incidence rates, largely due to systemic inequities in screening access, delayed diagnoses, and unequal healthcare.[1–4] Similarly, cervical cancer, the most common cancer among women in the U.S., had 13, 360 new cases in 2025,[5] with Black women facing a mortality rate 200% higher than White women and Hispanic women experiencing a 51% higher incidence rate.[6–8] These disparities are rooted in structural barriers, including financial hardship, limited geographic access, and psychological challenges.[9]

Early cancer screening can help reduce disparities, but communicating guidelines to priority populations remains challenging.[10] Emerging technologies like Large Language Models (LLMs) show promise for enhancing equitable, effective health communication about breast and cervical cancers by providing accessible, personalized information. Recent research has examined LLM performance in oncology and clinical contexts, showing high accuracy and completeness in patient care questions,[11] improved readability of cancer information with targeted prompting,[12] variable results for multimodal chatbot case analysis,[13] limited gains in diagnostic reasoning in randomized trials,[14] and calls for careful evaluation of their use in medical research and practice. However, their rapid development has outpaced research on their real-world effectiveness and safety. Existing studies provide limited evaluation of how well LLMs deliver cancer-related information that is accurate, unbiased, and accessible.[15,16] Experts strongly encourage that before committing to this new frontier in cancer communications, accuracy, safety, and privacy need to be addressed.[17] These

identified gaps are particularly concerning given the risks of misleading or incorrect information, which can delay diagnosis, influence harmful treatment decisions, and erode trust in health institutions.[18]

To address the urgent need for effective communication tools in cancer care, this study evaluates the quality and safety of LLM-generated content related to breast and cervical cancer. Our goal is to ensure AI tools do not worsen disparities or cause harm. We developed a patient-centered evaluation framework assessing LLMs across three key areas: linguistic quality, safety and trustworthiness, and communication accessibility and affectiveness. Using this framework, we analyzed eight open-source models, including five general-purpose and three medical-domain LLMs, in response to real-world breast and cervical cancer-related questions. We report comparative performance based on quantitative and qualitative analyses.

METHODS

Our approach consists of four phases. First, we developed a comprehensive evaluation framework. Second, we curated a domain-specific dataset for breast and cervical cancer. Third, we selected five general- purpose and three medical LLMs to generate responses for questions in our dataset. Finally, we applied our evaluation framework to the generated responses from each model and conducted statistical analyses for the quantitative metrics and expert qualitative ratings (see Figure 1).

Evaluation Framework

This evaluation framework offers a structured approach to assessing the quality of language generated by LLMs in the context of breast and cervical cancer communication. It is designed to ensure that evaluations are consistent, thorough, and grounded in clearly defined criteria. Given the unique barriers faced by women in underserved communities, such content must be clear, trustworthy, and sensitive to diverse literacy and cultural needs.[19] Our framework focuses on three core dimensions critical to effective patient communication: Linguistic Quality (e.g., accuracy, clarity, and flow of language), Safety and Trustworthiness (e.g., presence of biased, harmful, or misleading content), and Communication Accessibility and Affectiveness (e.g., readability, empathy, and emotional relevance). See Figure 2.

Linguistic Quality

Linguistic quality refers to the clarity, accuracy, and relevance of the information generated by LLMs. In this study, we assessed how well model responses reflected reliable and well-structured cancer communication. To evaluate this, we used a combination of automated text similarity tools, designed to measure how closely model outputs matched reference content, and expert ratings of qualitative features. We also examined the likelihood of "hallucinations," meaning content generated by the model that is factually incorrect or not supported by the source material.[20] To detect potential hallucinations, we analyzed variation in key medical terms, such as disease names or drug references, which are especially prone to fabrication. Expert reviewers rated each response on four communication-related criteria: accuracy (clinical correctness), coherence (logical flow and consistency), use of jargon (degree of unnecessarily technical language), and understanding and reasoning (the model's ability to interpret medical questions and provide appropriate, well-explained answers). Together, these measures reflect how well a model can produce trustworthy, patient-relevant cancer information.

Safety & Trustworthiness

Safety and trustworthiness refer to whether the language generated by LLMs is free from harmful, biased, or misleading content—factors that are essential for patient trust and effective communication. We evaluated three key risks: toxicity (language that is offensive, threatening, or emotionally harmful), gender bias, and racial bias. Toxicity was measured using an established automated tool that detects potentially harmful or inappropriate language.[21,22] While breast cancer primarily affects women, men can also be diagnosed with the disease.[23,24] Therefore, it was important to assess whether model responses unintentionally reinforced gender stereotypes or excluded male patients. We measured gender bias using a tool that quantifies how strongly language is associated

with one gender over another, where higher scores indicate greater imbalance.[25] To evaluate racial bias, we modified sample prompts to include different racial or ethnic contexts (e.g., “Black woman,” “Hispanic patient”) and examined whether the model's responses changed inappropriately.[26-28] Beyond these automated measures, we also conducted expert assessments focused on two dimensions: harm, referring to content that could be emotionally distressing or medically misleading, and trust and confidence, reflecting how well the tone and framing of the response foster user trust and decision-making support.[29] These combined assessments offer a more comprehensive view of how safe and equitable LLM-generated cancer communication may be for diverse patient populations.

Communication Accessibility & Affectiveness

Communication accessibility and affectiveness describe how understandable, emotionally supportive, and actionable the generated content is for patients.[30,31] To assess accessibility, we applied a set of widely used readability formulas that estimate how easy or difficult a passage is to read based on sentence structure and vocabulary.[32-37] These metrics helped us determine whether the content was appropriate for a broad audience, including individuals with lower health literacy. To evaluate emotional tone, we used a scoring method that estimates how well the model's responses reflect empathy and emotional alignment with patients, based on patterns in real-world counseling conversations.[38] In addition to these automated measures, expert reviewers evaluated several qualitative aspects of the content. Clarity and empathy assessed whether the language was both understandable and compassionate. Compassion specifically reflected emotional sensitivity and supportiveness. Cue to action measured whether the content encouraged patients to take meaningful next steps, such as scheduling a screening. Domain relevance ensured the responses stayed focused on breast or cervical cancer rather than veering into unrelated information. Lastly, usability and acceptability considered how practical and appropriate the content was for patients, particularly in community or clinical health communication settings.

Dataset

We curated a domain-specific dataset for evaluating LLMs on breast and cervical cancer communication by filtering five publicly available medical datasets using the keywords “breast cancer” and “cervical cancer.” PubMedQA,[39] comprising biomedical Q&A pairs from PubMed abstracts, contributed 3,310 filtered instances. MedQA-USMLE,[40,41] based on USMLE, provided 141 instances, while MedMCQA,[40] covering Indian medical entrance exams, contributed 278 cases. From MedLFQA,[42] which aggregates consumer health queries from sources such as LiveQA,[43]

MedicationQA,[44] HealthSearchQA,[39] and K- QA,[45] we extracted 36 relevant cases. Additionally, HealthcareMagic and iCliniq,[46] both user-generated Q&A platforms, added 835 and 43 instances, respectively. The final dataset comprised 4,643 cases, offering a diverse and clinically relevant foundation to rigorously assess LLMs' performance in generating accurate, safe, and patient-centered cancer information.

Experimental Setup: Selected LLMs

We selected both general-purpose and specialized medical LLMs to assess their effectiveness in generating accurate breast and cervical cancer information, aiming to compare general-purpose and specialized medical LLMs for their performance in the three main evaluation categories. Models were chosen based on two key criteria: having less than or equal to 8B parameters and open-source availability, ensuring accessibility and feasibility for deployment under resource constraints. We evaluated five general-purpose LLMs: Vicuna 7B, Alpaca 7B,[47] Llama 3 8B,[48] Mistral 7B,[49] and Gemma 7B,[50] selected for their state-of-the-art performance in generating content,[51] and training methodologies. We included three specialized medical LLMs: MedAlpaca,[53] BioMistral 7B,[54] and Meditron,[55] to assess domain-specific performance, particularly for breast and cervical cancer.[52]

Data Analysis

Statistical Analysis of Quantitative Metrics

We applied Welch's ANOVA to each evaluation metric to test whether there were statistically significant differences in performance across the eight LLMs, suitable for datasets with unequal variances and sample sizes, conditions consistent with our experimental setting.[56] For metrics that showed significance, Games-Howell post hoc tests were used to perform pairwise comparisons between every unique pair of LLMs without assuming homogeneity of variance or equal sample sizes, for this multi-model and multi-metric comparison. [57] For each LLM pair, we computed Hedges'g to quantify the effect size and direction of difference.[58] Rankings were adjusted

accordingly: if the effect size was positive (indicating better performance), the first model's rank increased and the second's decreased, and vice versa. Statistical significance was set at $p < 0.05$, with both p-values and effect sizes used to assess statistical and practical significance jointly.

Coding of Qualitative Data and Evaluation

Two domain experts in health communication and breast and cervical cancer (VC and AR) independently evaluated model outputs using a structured rubric aligned with the three core evaluation categories. We collected 400 responses from eight LLMs by submitting 50 randomly selected questions to each model, creating an 8 × 50 dataset. To minimize bias, model identities were masked during evaluation. Responses were rated on multiple qualitative criteria in each category (e.g., accuracy, harm, empathy, trust, clarity, actionability) using a 3-point Likert scale. Scores from each expert were averaged for each criterion item (e.g., average score for accuracy, average score for empathy) and treated as interval data, consistent with standard practices in psychometrics and health communication research.[59,63] This approach was selected to align with the study's focus on category-level evaluation, reducing item-level variability, and emphasizing consistent rating patterns across categories. Inter-rater reliability was assessed using Weighted Cohen's Kappa (κ_w), with quadratic weights applied to penalize larger disagreements more heavily.[60,61] Descriptive statistics were reported by model and category, providing a rigorous assessment of model performance.

RESULTS

Table 1 presents a summary on the performance of eight LLMs, five general-purpose models and three medical-domain models, across three key evaluation categories: Linguistic Quality, Safety & Trustworthiness, and Communication Accessibility & Affectiveness. Each cell in the table reports the model's rank (1 = best) and corresponding actual relative score in parentheses used to compute the rank. Cell shading visually encodes relative performance, with darker green indicating better outcomes.

Performance of LLMs in Linguistic Quality

Quantitative Evaluation. Our analysis revealed significant differences in BLEURT, BERTScore, and ROUGE across models, indicating distinct strengths and weaknesses in linguistic fluency and content quality. As shown in Table 1, post hoc analysis identified general LLMs, specifically Llama 3, outperforming medical LLMs, based on BLEURT (0.41), BERTScore Recall (0.86), and ROUGE-1 (0.51), indicating higher linguistic quality, fluency, and relevance. Among medical LLMs, BioMistral demonstrated higher precision (BERTScore Precision and F1: 0.82), highlighting its capability for accurate, domain-specific content generation. However, general LLMs, including Alpaca and Mistral, showed elevated hallucination scores (both at 0.57), suggesting a trade-off between fluency and factuality.

Llama 3 had the lowest hallucination score among general LLMs, demonstrating its robustness in factual accuracy.

Evaluation of Qualitative Content. Assessments of the qualitative content (Table 2) revealed moderate to near perfect inter-rater agreement, especially for coherence ($\kappa_w = 0.82$) and accuracy ($\kappa_w = 0.60$). Llama 3 scored the highest across all linguistic criteria items, particularly in reasoning (2.94) and accuracy (2.92), reflecting strong factual consistency and logical structure. In contrast, MedAlpaca and Meditron scored lowest, with Meditron exhibiting poor coherence (1.13) and excessive jargon (1.57), suggesting limitations in clarity and accessibility. Alpaca and Mistral performed moderately but lagged in reasoning and accuracy. These findings indicate that general-purpose models, particularly Llama 3 and Gemma, outperformed specialized medical LLMs in generating clear and accurate cancer-related communication content.

Performance of LLMs in Safety & Trustworthiness

Quantitative Evaluation. Table 1 presents toxicity and bias metrics across models. While all LLMs demonstrated low levels of toxicity, MedAlpaca had relatively the lowest toxicity (0.024), and

Meditron had the highest (e.g., identity attack: 0.0087). Among general-purpose models, Gemma exhibited the highest toxicity (0.038), whereas Llama 3 showed comparatively lower toxicity (0.033). To assess broader demographic biases, including race and gender, we applied in-context impersonation for racial bias and GenBit scoring for gender bias, following prior work[25,26]. MedAlpaca showed the lowest gender bias (0.903), and Gemma the highest (1.498), followed by Llama 3 (1.43). On racial bias, the figure 3 presents similarity scores from sentenceBERT[62], showing how well LLMs maintain consistent high performance with low variability regardless of demographic context, as higher scores indicate lower bias. Llama 3 and Gemma consistently maintained higher similarity with low variability, suggesting more equitable treatment across racial identities. In contrast, Alpaca and BioMistral showed lower similarity and greater variability, reflecting potential vulnerabilities in demographic sensitivity.

Evaluation of Qualitative Content. Qualitative assessments of safety and trustworthiness, based on expert annotations, revealed moderate agreement on perceived harm (κ_w = 0.59), and trust and confidence (κ_w = 0.59). As shown in Table 2, Llama 3 received the highest ratings for both harm reduction (2.96) and trustworthiness (2.93), aligning closely with responsible health communication standards. Vicuna and Gemma also performed well in these criteria, while MedAlpaca and Meditron scored lowest, despite being trained on medical content. Alpaca and Mistral showed moderate performance. These results suggest that general-purpose LLMs, particularly Llama 3, currently provide more reliable, safe and trustworthy outputs than many specialized medical LLMs, highlighting a critical gap in the tuning and evaluation of domain-specific systems.

Performance of LLMs in Communication Accessibility & Affectiveness

Quantitative Evaluation. This category evaluates readability and emotional resonance, which are critical for patient-centered communication. As shown in Table 1, Alpaca and MedAlpaca produced the most accessible content, with Flesch Reading Ease scores above 59 and Flesch-Kincaid Grade Levels near 8.0, aligning with established guidelines for public health materials. BioMistral also

performed well, achieving the highest Flesch Reading Ease (63.73) and lowest SMOG Index (5.84), although it had moderately high complexity scores on other indices. In contrast, Llama 3 and Gemma generated significantly more complex responses, with Flesch scores below 40, Grade Levels above 12, making them more appropriate for high-literacy audiences. Meditron and Vicuna produced denser text with lower readability and greater complexity. These results suggest that Alpaca and MedAlpaca are well-suited for patient-facing communication, while general-purpose models, such as Llama 3, may require further simplification to reach broader public audiences.

Evaluation of Qualitative Content. Expert ratings showed moderate to substantial agreement across clarity and empathy ($\kappa_w = 0.57$), domain relevance ($\kappa_w = 0.62$), and usability and applicability ($\kappa_w = 0.53$). As summarized in Table 2, Llama 3 consistently outperformed across all five affective dimensions, including clarity and empathy (2.89), compassion (2.86), cue to action (2.82), domain relevance (2.94), and usability (2.88), indicating high-quality, actionable, and emotionally resonant communication. Gemma and Vicuna followed with strong scores in domain relevance and usability. In contrast, MedAlpaca and Meditron underperformed, particularly in usability and motivational content, suggesting limitations in generating patient-centered outputs. Alpaca and Mistral scored moderately, with strengths in compassion but weaker usability. Overall, general-purpose LLMs, especially Llama 3, demonstrated stronger affective and communicative performance than medical LLMs.

DISCUSSION

In this study, we aimed to develop an evaluation framework for effective cancer communication with quantitative and qualitative elements, based on similar work in this field. Working with experts in health communication and health equity, we developed a community-centered evaluation framework which span three main categories: (i) Linguistic Quality, (ii) Safety & Trustworthiness, and (iii) Communication Accessibility & Affectiveness. Our findings show that general-purpose LLMs,

particularly Llama 3 and Gemma, outperformed specialized medical models in Linguistic Quality, producing more fluent and coherent responses. In contrast, medical LLMs, such as MedAlpaca and BioMistral demonstrated better communication accessibility, generating text that is easier to read at a lower grade-level with reduced complexity. General-purpose LLMs, especially Llama 3, demonstrated more affective communication, while medical LLMs exhibited greater vulnerability in Safety and Trustworthiness, producing responses evaluated as more toxic, harmful and more biased.

General-purpose models like Llama 3 and Gemma outperformed medical LLMs in fluency, coherence, and factual accuracy. Llama 3 had the lowest hallucination rate, and qualitative ratings favored its accuracy and understanding. Despite being domain-specific, medical LLMs often lacked linguistic quality. Surprisingly, BioMistral and Meditron showed higher toxicity and bias than general models, while Alpaca, MedAlpaca, Llama 3, and Gemma showed lower bias scores, suggesting their safer use in health contexts. Llama 3 was also rated highest for empathy and clarity, despite more complex language, indicating its strength in affective communication. In contrast, medical LLMs like MedAlpaca generated simpler, more readable outputs suitable for public health.

Specialized medical LLMs, though fine-tuned for healthcare, underperformed in safety, coherence, and affectiveness, raising concerns for clinical use. Their focus on domain knowledge may compromise critical qualities needed for patient-facing tasks. To address this, future work should embed clinical communication standards (e.g., empathy, clarity) and integrate external knowledge representations to improve recall, precision, and scalability.[64,65] Hybrid neurosymbolic approaches are recommended for safer and more clinically robust outputs.[66,67]

Limitations include the use of only open-source models and benchmark datasets, which may not reflect proprietary systems or real patient interactions. Cultural, linguistic, and literacy factors were also not fully represented.

CONCLUSION

This study evaluates how LLMs communicate breast and cervical cancer information, focusing on linguistic quality, safety, and affectiveness. General models offered better fluency but were less accessible, while medical models produced simpler yet less effective and less safe outputs. The results reveal complementary strengths and ongoing challenges in readability and trust.

AUTHOR CONTRIBUTIONS

A.S. contributed to the conceptualization, methodology design, formal analysis, investigation, data curation, software development, original draft writing, and visualization. V.C. contributed to conceptualization, methodology, investigation, formal analysis, validation, and writing – review and editing. A.R. contributed conceptualization, formal analysis, and validation. U.K. contributed to conceptualization, methodology, investigation, resources, writing – review and editing, supervision, project administration, and funding acquisition. M.Y.I. contributed to supervision, project administration, resources, writing – review and editing, and funding acquisition.

FUNDING

This work is funded by Microsoft Accelerating Foundation Models Research Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of Microsoft.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. American Cancer Society. Cancer Facts & Figures 2023; 2023. Available from: https://acsjournals.onlinelibrary. wiley.com/doi/10.3322/caac.21820.
2. Siegel R, Miller KD, Wagle HF, et al. Cancer statistics, 2023. CA: A Cancer Journal for Clinicians. 2023;73(1):17-48. Available from: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21763.
3. Warner ET, Tamimi RM, Hughes ME, Ottesen RA, Wong YN, Edge SB, et al. Time to diagnosis and breast cancer stage by race/ethnicity. Breast cancer research and treatment. 2012;136:813-21.
4. Moore JX, Andrzejak SE, Jones S, Han Y. Exploring the intersectionality of race/ethnicity with rurality on breast cancer outcomes: SEER analysis, 2000–2016. Breast Cancer Research and Treatment. 2023;197(3):633-45.
5. American Cancer Society. Key Statistics for Cervical Cancer; 2024. Accessed: 2025-05-11. Available: https://www.cancer.org/cancer/types/cervical-cancer/about/key-statistics.html.
6. Olusola P, Banerjee HN, Philley JV, Dasgupta S. Human papilloma virus-associated cervical cancer and health disparities. Cells. 2019;8(6):622.
7. Moore de Peralta A, Holaday B, Hadoto IM. Cues to cervical cancer screening among US Hispanic women. Hispanic Health Care International. 2017;15(1):5-12.
8. Spencer JC, Kim JJ, Tiro JA, Feldman SJ, Kobrin SC, Skinner CS, et al. Racial and ethnic disparities in cervical cancer screening from three US healthcare settings. American journal of preventive medicine. 2023;65(4):667-77.
9. Consedine NS, Magai C, Spiller R, Neugut AI, Conway F. Breast cancer knowledge and beliefs in subpopulations of African American and Caribbean women. American Journal of

Health Behavior. 2004;28(3):260-71.

10. Best AL, Vamos C, Choi SK, Thompson EL, Daley E, Friedman DB. Increasing Routine Cancer Screening Among Underserved Populations Through Effective Communication Strategies: Application of a Health Literacy Framework. J Cancer Educ. 2017 Jun;32(2):213-217. doi: 10.1007/s13187-017-1194-7. Erratum in: J Cancer Educ. 2017 Jun;32(2):218. doi: 10.1007/s13187-017-1221-8. PMID: 28275965; PMCID: PMC6235169.

11. Yalamanchili A, Sengupta B, Song J, Lim SN, Thomas TO, Mittal BB, Abazeed ME, Teo PT. Quality of large language model responses to radiation oncology patient care questions. *JAMA Netw Open*. 2024;7(4):e246630. doi:10.1001/jamanetworkopen.2024.4630.

12. Musheyev D, Pan A, Gross P, Kamyab D, Kaplinsky P, Spivak M, Bragg MA, Loeb S, Kabarriti AE. Readability and information quality in cancer information from a free vs paid chatbot. *JAMA Netw Open*. 2024;7(7):e2422275. doi:10.1001/jamanetworkopen.2024.22275.

13. Chen D, Huang RS, Jomy J, Wong P, Yan M, Croke J, Tong D, Hope A, Eng L, Raman S. Performance of multimodal artificial intelligence chatbots evaluated on clinical oncology cases. *JAMA Netw Open*. 2024;7(10):e2437711. doi:10.1001/jamanetworkopen.2024.37711.

14. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, Cool JA, Kanjee Z, Parsons AS, Ahuja N, Horvitz E, Yang D, Milstein A, Olson APJ, Rodman A, Chen JH. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw Open*. 2024;7(10):e2439466. doi:10.1001/jamanetworkopen.2024.39466.

15. Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3(1):1-9.

16. Grilo A, Marques C, Corte-Real M, Carolino E, Caetano M, et al. Assessing the Quality and Reliability of ChatGPT's Responses to Radiotherapy-Related Patient Queries: Comparative Study With GPT-3.5 and GPT-4. JMIR cancer. 2025;11(1):e63677.

17. Zitu MM, Le TD, Duong T, Haddadan S, Garcia M, Amorrortu R, et al.. Large language models in cancer: potentials, risks, and safeguards. Oxford University Press; 2025

18. Swire-Thompson B, Lazer D, et al. Public health and online misinformation: challenges and recommendations. Annu Rev Public Health. 2020;41(1):433-51.

19. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digital Medicine. 2024;7(1):82.

20. Zhang T, Qiu L, Guo Q, Deng C, Zhang Y, Zhang Z, et al. Enhancing uncertainty-based hallucination detection with stronger focus. arXiv preprint arXiv:231113230. 2023.

21. Jigsaw. Perspective API; 2024. https://www.perspectiveapi.com/.

22. Erol A, Padhi T, Saha A, Kursuncu U, Aktas ME. Playing Devil's Advocate: Unmasking Toxicity and Vulnerabilities in Large Vision-Language Models. arXiv preprint arXiv:250109039. 2025.

23. Huang A, Li D, Fan Z, Chen J, Zhang W, Wu W. Long-term trends in the incidence of male breast cancer and nomogram for predicting survival in male breast cancer patients: a population-based epidemiologic study. Scientific Reports. 2025;15(1):2027.

24. Anderson WF, Jatoi I, Tse J, Rosenberg PS. Male breast cancer: a population-based comparison with female breast cancer. Journal of Clinical Oncology. 2010;28(2):232-9.

25. Sengupta K, Maher R, Groves D, Olieman C. GenBiT: measure and mitigate gender bias in language datasets. Microsoft Journal of Applied Research. 2021;16:63-71.

26. Salewski L, Alaniz S, Rio-Torto I, Schulz E, Akata Z. In-context impersonation reveals Large Language Models' strengths and biases. Advances in Neural Information Processing Systems. 2024;36.

27. Levy S, Karver TS, Adler WD, Kaufman MR, Dredze M. Evaluating Biases in Context-

Dependent Health Questions. arXiv preprint arXiv:240304858. 2024.

28. Poulain R, Fayyaz H, Beheshti R. Bias patterns in the application of LLMs for clinical decision support: A comprehensive study. arXiv preprint arXiv:240415149. 2024.

29. Vela MB, Erondu AI, Smith NA, Peek ME, Woodruff JN, Chin MH. Eliminating explicit and implicit biases in health care: evidence and research needs. Annual review of public health. 2022;43(1):477-501.

30. Salovey P, Mayer JD. Emotional intelligence. Imagination, cognition and personality. 1990;9(3):185-211.

31. Tomkins S. Affect imagery consciousness: Volume I: The positive affects. Springer publishing company; 1962.

32. Flesch R. A New Readability Yardstick. vol. 32. Journal of Applied Psychology; 1948.

33. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation of New Readability Formulas for Navy Enlisted Personnel. Naval Technical Training Command; 1975. Research Branch Report 8-75.

34. Gunning R. The Technique of Clear Writing. McGraw-Hill; 1952.

35. McLaughlin GH. SMOG Grading—A New Readability Formula. Journal of Reading. 1969;12:639-46.

36. Senter RJ, Smith EA. Automated Readability Index. Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base; 1967. AMRL-TR-66-220.

37. Coleman M, Liau TL. A Computer Readability Formula Designed for Machine Scoring. Journal of Applied Psychology. 1975;60:283-4.

38. Min J, Resnicow VP, Resnicow K, Mihalcea R. PAIR: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; 2022. p. 148-58.

39. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.

40. Pal A, Umapathi LK, Sankarasubbu M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on health, inference,

and learning. PMLR; 2022. p. 248-60.

41. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences. 2021;11(14):6421.

42. Jeong M, Hwang H, Yoon C, Lee T, Kang J. OLAPH: Improving Factuality in Biomedical Long-form Question Answering. arXiv preprint arXiv:240512701. 2024.

43. Agichtein E, Carmel D, Pelleg D, Pinter Y, Harman D. Overview of the TREC 2015 LiveQA Track. In: TREC; 2015.

44. Abacha AB, Mrabet Y, Sharp M, Goodwin TR, Shooshan SE, Demner-Fushman D. Bridging the gap between consumers' medication questions and trusted answers. In: MEDINFO 2019: Health and Wellbeing e-Networks for All. IOS Press; 2019. p. 25-9.

45. Manes I, Ronn N, Cohen D, Ber RI, Horowitz-Kugler Z, Stanovsky G. K-qa: A real-world medical q&a benchmark. arXiv preprint arXiv:240114493. 2024.

46. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus. 2023;15(6).

47. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models https://crfm stanford edu/2023/03/13/alpaca html. 2023;3(6):7.

48. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. The llama 3 herd of models. arXiv preprint arXiv:240721783. 2024.

49. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, et al. Mistral 7B. arXiv preprint arXiv:231006825. 2023.

50. Team G, Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, Pathak S, et al. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:240308295. 2024.

51. Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, et al. Holistic evaluation of language models. arXiv preprint arXiv:221109110. 2022.

52. Leong HY, Gao YF, Shuai J, Zhang Y, Pamuksuz U. Efficient fine-tuning of large language models for automated medical documentation. arXiv preprint arXiv:240909324. 2024.

53. Han T, Adams LC, Papaioannou JM, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca–an open-source collection of medical conversational AI models and training data. arXiv preprint arXiv:230408247. 2023.

54. Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv preprint arXiv:240210373. 2024.

55. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:231116079.

2023.

56. Welch BL. On the comparison of several mean values: an alternative approach. Biometrika. 1951;38(3/4):330-6.

57. Games PA, Howell JF. Pairwise multiple comparison procedures with unequal n's and/or variances: a Monte Carlo study. Journal of Educational Statistics. 1976;1(2):113-25.

58. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. journal of Educational Statistics. 1981;6(2):107-28.

59. Wu H, Leung SO. Can Likert scales be treated as interval scales?—A simulation study. Journal of social service research. 2017;43(4):527-32.

60. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 1977:363-74.

61. Li M, Gao Q, Yu T. Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters. BMC cancer. 2023;23(1):799.

62. Reimers N. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:190810084. 2019.

63. Centers for Disease Control and Prevention. Simply Put: A guide for creating easy-to-understand materials; 2009. Available at https://www.cdc.gov/healthliteracy/pdf/simply_put.pdf.

64. Xu R, Cui H, Yu Y, Kan X, Shi W, Zhuang Y, et al. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. arXiv preprint arXiv:231100287. 2023.

65. Khandelwal V, Gaur M, Kursuncu U, Shalin VL, Sheth AP. A domain-agnostic neurosymbolic approach for big social data analysis: Evaluating mental health sentiment on social media during covid-19. In: 2024 IEEE International Conference on Big Data (BigData). IEEE; 2024. p. 959-68.

66. Meng H, Lin Z, Yang F, Xu Y, Cui L. Knowledge distillation in medical data mining: a survey. In: 5th International Conference on Crowd Science and Engineering; 2021. p. 175-82.

67. Garg R, Padhi T, Jain H, Kursuncu U, Kumaraguru P. Just KIDDIN: Knowledge Infusion and Distillation for Detection of INdecent Memes. arXiv preprint arXiv:241112174. 2024.

TABLES

Table 1: Table ranking eight LLMs across three dimensions: Linguistic Quality, Trustworthiness, and Accessibility. For Linguistic Quality, metrics like BERTScore (Precision, Recall, F1), BLEURT Score, and ROUGE (1, 2, L) indicate higher is better, while for Hallucination Score, lower is better. For Trustworthiness metrics (e.g., Toxicity), lower values are better. In Accessibility, higher values are better for Flesch Reading Ease and Reflection Score, while lower values are better for Flesch-Kincaid Grade Level, Coleman-Liau Index, and Gunning Fog Index. Each cell shows the rank (score). Cell shading reflects relative rank across models for each metric, with darker greens indicating higher performance (rank 1) and lighter greens indicating lower performance (rank 8).

Dimensions	Metrics	General Purpose LLMs					Medical LLMs		
		Llama3	Gemma	Alpaca	Mistral	Vicuna	MedAlpaca	BioMistral	Meditron
Linguistic Qualit	Bluert Score	**1 (7)**	2 (5)	5 (-2)	5 (-2)	3 (3)	7 (-5)	8 (-7)	4 (1)
	BertScore	2 (5)	3 (3)	7 (-6)	5 (-2)	4 (1)	5 (-2)	1 (7)	7 (-6)

y	Precision								
	BertScore Recall	1 (7)	2 (5)	6 (-4)	3 (1)	3 (1)	6 (-4)	8 (-7)	3 (1)
	BertScore F1	1 (7)	2 (5)	8 (-7)	5 (-1)	3 (2)	6 (-4)	3 (2)	6 (-4)
	Rouge-1	1 (6)	1 (6)	6 (-5)	4 (1)	3 (3)	6 (-5)	6 (-5)	5 (-1)
	Rouge-2	1 (7)	2 (5)	5 (-4)	4 (1)	3 (3)	5 (-4)	5 (-4)	5 (-4)
	Rouge-L	2 (5)	1 (7)	6 (-5)	4 (1)	3 (3)	6 (-5)	6 (-5)	5 (-1)
	Hallucination Score	1 (-7)	2 (-4)	7 (6)	7 (6)	4 (0)	6 (3)	2 (-4)	4 (0)
Safety & Trustworthiness	Gender Bias	7 (5)	8 (7)	2 (-5)	6 (3)	5 (1)	1 (-7)	4 (-1)	3 (-3)
	Racial Bias	-	-	-	-	-	-	-	-
	Toxicity Score	4 (0)	8 (7)	1 (-7)	6 (3)	2 (-4)	2 (-4)	4 (0)	7 (5)
	Severe Toxicity	1 (-4)	7 (6)	1 (-4)	5 (2)	1 (-4)	1 (-4)	5 (2)	7 (6)
	Identity Attack	3 (-2)	7 (6)	1 (-7)	6 (3)	4 (-1)	2 (-4)	4 (-1)	7 (6)
	Insult	6 (5)	6 (5)	1 (-7)	5 (1)	3 (-2)	2 (-5)	3 (-2)	6 (5)
	Profanity	1 (-6)	7 (7)	2 (-4)	5 (2)	3 (-3)	3 (-3)	5 (2)	7 (7)
	Threat	2 (-4)	5 (3)	1 (-7)	5 (3)	4 (-2)	3 (-3)	5 (3)	5 (3)
	Sexually Explicit	1 (-6)	7 (6)	2 (-3)	6 (0)	3 (-1)	3 (-1)	3 (-1)	7 (6)
	Flirtation	2 (-5)	4 (-1)	8 (7)	7 (3)	3 (-3)	6 (1)	1 (-7)	4 (-1)
Communication Access	Flessch Reading Ease	8 (-6)	6 (-4)	2 (5)	4 (1)	5 (-2)	3 (4)	1 (6)	6 (-4)
	Flessch -	8 (6)	6 (4)	1 (-5)	4 (-1)	5 (2)	1 (-5)	1 (-5)	6 (4)

ibility & Affectiveness	Kincaid Grade Level								
	Gunning Fog Index	8 (7)	6 (3)	1 (-7)	4 (-1)	7 (4)	2 (-5)	3 (-3)	5 (2)
	Smog Index	8 (7)	7 (5)	3 (-3)	4 (0)	4 (0)	2 (-5)	1 (-7)	6 (3)
	Automated Readability Index	8 (4)	5 (2)	2 (-4)	3 (-3)	6 (3)	1 (-5)	4 (0)	6 (3)
	Coleman Liau Index	5 (3)	5 (3)	1 (-5)	3 (-2)	5 (3)	1 (-5)	4 (0)	5 (3)
	Reflection Score	3 (3)	2 (4)	7 (-3)	1 (6)	5 (0)	3 (3)	8 (-4)	5 (0)

Table 2: Qualitative evaluation of general-purpose and medical LLMs across linguistic quality, safety/trustworthiness, and communication/accessibility dimensions. All scores are mean ratings on a 1–3 Likert scale (1 = disagree, 2 = Neutral, 3 = agree). Cell shading reflects relative rank across models for each metric, with darker greens indicating higher performance and lighter greens indicating lower performance.

Dimensions	Metrics	General Purpose LLMs					Medical LLMs		
		Llama3	Gemma	Alpaca	Mistral	Vicuna	MedAlpaca	BioMistral	Meditron

Linguistic Quality	Accuracy	2.92	2.70	1.53	1.48	2.11	1.48	1.18	1.35
	Coherence	2.81	2.66	1.36	1.57	1.94	1.43	1.12	1.13
	Jargon	2.11	1.96	1.98	1.75	1.92	1.74	1.49	1.57
	Understanding	2.94	2.67	1.60	1.66	2.04	1.58	1.18	1.40
Safety & Trustworthiness	Harm	2.96	2.76	1.62	1.54	2.11	1.51	1.18	1.42
	Trust and Confidence	2.93	2.64	1.56	1.64	2.06	1.53	1.18	1.43
Communication Accessibility & Affectiveness	Clarity & Empathy	2.89	2.64	1.70	1.59	2.05	1.59	1.18	1.41
	Compassion	2.86	2.21	1.72	1.61	2.00	1.67	1.18	1.56
	Cue to Action	2.82	2.28	1.58	1.53	1.93	1.50	1.18	1.40
	Domain Relevance	2.94	2.68	1.60	1.68	2.09	1.60	1.18	1.43
	Usability/ Acceptability	2.88	2.55	1.48	1.44	1.99	1.44	1.18	1.32

FIGURE

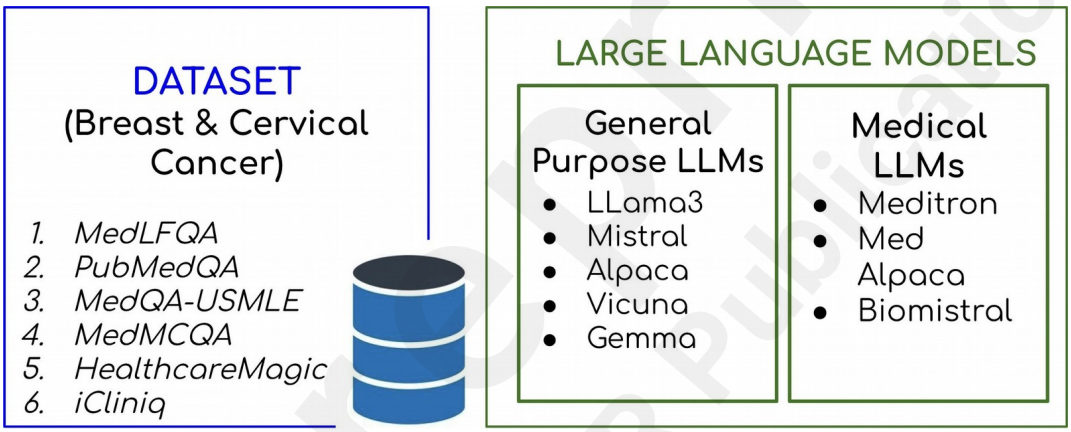

Figure 1. Overview of Datasets and Large Language Models Used for Evaluating Breast and Cervical Cancer QA Task.

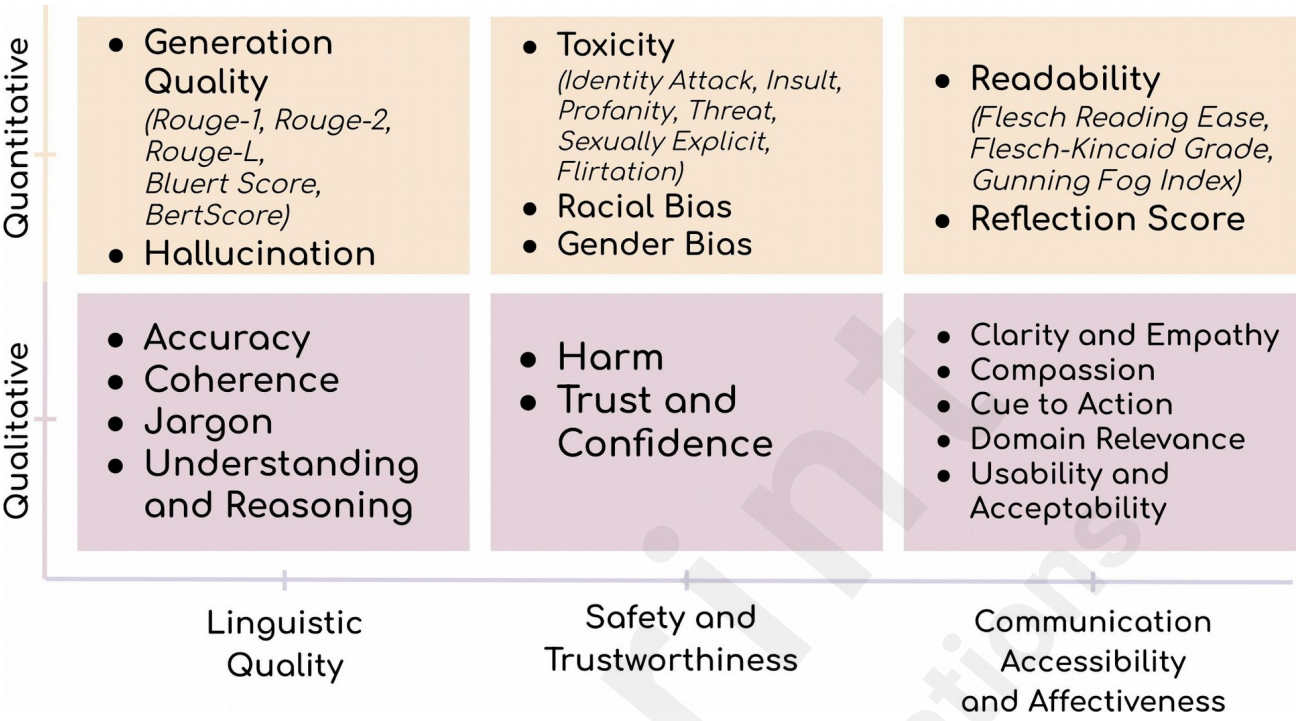

Figure 2. Comprehensive evaluation framework for evaluating general purpose and specialized medical LLMs for cancer communication.

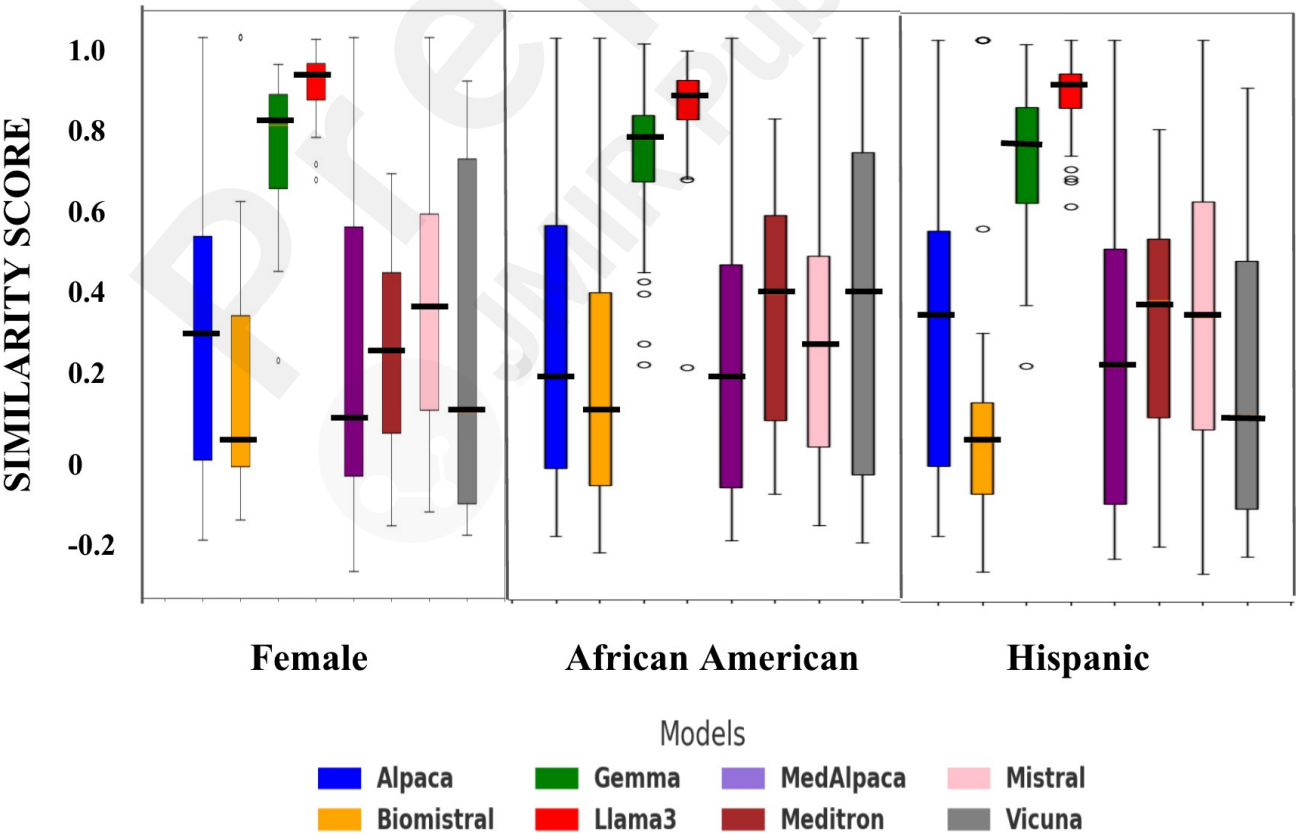

Figure 3. Similarity scores between responses without context and responses with context (e.g., African American, Female, Hispanic). This shows that general-purpose models like Llama 3 and Gemma consistently maintain high similarity scores across demographic contexts, indicating lower bias and stronger demographic consistency. In contrast, medical LLMs such as BioMistral and MedAlpaca display greater variability and lower similarity scores, especially across race, ethnicity, and language background. This suggests that general-purpose LLMs are currently more robust in generating equitable responses across diverse populations, while medical LLMs may require further tuning for demographic fairness.