

Beyond the Rubric: Cultural Misalignment in LLM Benchmarks for Sexual and Reproductive Health

Sumon Kanti Dey¹, Manvi S¹, Zeel Mehta³, Meet Shah³,
Unnati Agrawal¹, Suhani Jalota^{2,3}, Azra Ismail¹

¹Emory University, Atlanta, Georgia, USA,

²Hoover Institution, Stanford University, Stanford, California, USA,

³Myna Mahila Foundation, Mumbai, India,

Correspondence: {sumon.kanti.dey, azra.ismail}@emory.edu

Abstract

Large Language Models (LLMs) have been positioned as having the potential to expand access to health information in the Global South, yet their evaluation remains heavily dependent on benchmarks designed around Western norms. We present insights from a preliminary benchmarking exercise with a chatbot for sexual and reproductive health (SRH) for an underserved community in India. We evaluated using HealthBench, a benchmark for conversational health models by OpenAI (Arora et al., 2025). We extracted 637 SRH queries from the dataset and evaluated on the 330 single-turn conversations. Responses were evaluated using HealthBench’s rubric-based automated grader, which rated responses consistently low. However, qualitative analysis by trained annotators and public health experts revealed that many responses were actually culturally appropriate and medically accurate. We highlight recurring issues, particularly a Western bias, such as for legal framing and norms (e.g., breastfeeding in public), diet assumptions (e.g., fish safe to eat during pregnancy), and costs (e.g., insurance models). Our findings demonstrate the limitations of current benchmarks in capturing the effectiveness of systems built for different cultural and healthcare contexts. We argue for the development of culturally adaptive evaluation frameworks that meet quality standards while recognizing needs of diverse populations. The code is available at¹.

1 Introduction

SRH is a critical aspect of overall wellbeing, yet access to accurate and empathetic guidance remains uneven across geographies, due to deep-seated cultural taboos, poor sex education, and gaps in information access (Wahyuningsih et al., 2024; Shaw, 2009). Increasingly, health organizations are investing in chatbots powered by LLMs to facilitate accessible SRH health guidance. However,

the effectiveness of LLMs does not just depend on accuracy, but also on their ability to deliver culturally relevant and contextually appropriate responses (Deva et al., 2025; Andalibi and Bowen, 2022). Yet, evaluation of such chatbots, in SRH and beyond, remains a persistent challenge. Recent work, such as *HealthBench* (Arora et al., 2025), provides a large-scale dataset and rubrics to evaluate LLMs on health tasks. While *HealthBench* has been developed with a global network of healthcare providers and aims to support evaluation of generalized healthcare chatbots, we find that it remains grounded in Western guidelines and assumptions.

In this paper, we present a preliminary evaluation of *Myna Bolo*, an LLM-based chatbot developed by the Myna Mahila Foundation (Myna), a Mumbai (India)-based Non-Governmental Organization (NGO) focused on women’s health and empowerment. We partnered with Myna to evaluate *Myna Bolo* that aims to provide localized and medically accurate SRH information for women from an underserved community in Mumbai, India. The system runs on WhatsApp, combines retrieval-augmented generation (RAG) with intent detection, and includes a human-in-the-loop option for escalation to experts. A screenshot of the interface is in Appendix A. To benchmark the performance of *Myna Bolo*, we turned to *HealthBench*, extracting SRH queries (n=637). As this is a preliminary study, we restrict our analysis to the 330 single-turn questions in the dataset. Through our qualitative analysis, we observed a systematic mismatch—culturally and regionally-appropriate responses grounded in an underserved Indian context were being scored as incorrect.

Through this study, we argue that while large benchmarks offer a standardized framework for evaluating health chatbots, they may overlook culture- and region-sensitivity. They also tend to be designed from a clinician perspective, and can miss a critical human-centered perspective that

¹<https://github.com/Sumon/healthbench-srh-eval/>

meets the needs of users. Our contributions are twofold: (1) We analyze how HealthBench (HB) rubrics, designed around Western norms, can penalize culturally-grounded SRH responses. (2) We discuss implications for evaluating LLMs in global health, highlighting the need for culturally adaptive benchmarks, particularly in low-resource settings.

2 Related Work

In recent years, several benchmarks for evaluating health LLMs have emerged, including MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), CareQA (Arias-Duart et al., 2025), and MedHELM (Bedi et al., 2025). Beyond these, a rapidly expanding set of evaluation metrics has been proposed, such as QUEST (Tam et al., 2024), Med-HALT (Pal et al., 2023), CSEDB (Wang et al., 2025), CRAFT-MD (Johri et al., 2025), AMIE (Tu et al., 2024). While these focus on domain-specific reasoning and safety of model-generated responses, they remain limited in scope for assessing real-world conversational systems. Among these efforts, HealthBench (Arora et al., 2025) has emerged as one of the most comprehensive frameworks that covers a wide range of health domains and enables systematic scoring of LLM model outputs. It also covers SRH, missing in many other benchmarks.

However, health information is highly context-dependent, shaped by medical practice, cultural norms, and resource availability (Brashers et al., 2002). Many SRH chatbots operate in low-resource languages and handle privacy-sensitive queries, where stigma and confidentiality are critical (e.g., SnehAI (Wang et al., 2022), AdolescentBot (Rahman et al., 2021), Nurse Nisa (McMahon et al., 2023)). Prior work in global health and medical NLP has noted that benchmarks developed in one cultural setting may not transfer directly to others (Hershcovich et al., 2022). For example, dietary advice, contraceptive methods, and even health-seeking vary significantly between regions. Evaluations that fail to recognize these differences risk undervaluing responses that are accurate and useful in local contexts (Nimo et al., 2025; Deva et al., 2025; Mutisya et al., 2025).

Additionally, many automatic grading systems rely on exact phrasing matches, disadvantaging culturally valid or concise responses that guide users correctly (Abd-Alrazaq et al., 2020; Abbasian et al., 2024). Our study focuses on HealthBench as a case study, analyzing gaps with contextual re-

quirements, to derive implications for health benchmarks broadly.

3 Data and Evaluation Setup

Our study offers a qualitative analysis of cultural misalignment in rubrics for queries on sexual and reproductive health in an LLM benchmark, specifically HealthBench. HealthBench (HB) is a physician-curated benchmark developed by OpenAI with 5,000 single- and multi-turn, clinically realistic conversations globally for evaluating conversational health models (Arora et al., 2025). For this study, we focused on SRH queries within HB. Using an LLM classifier (detailed prompt in Appendix B), we extracted 637 SRH queries. To mitigate potential selection bias, two human reviewers independently reviewed and verified all extracted queries to ensure correctness and SRH relevance. This includes 330 single-turn (51.8%) and 307 multi-turn (48.2%) conversations. For the preliminary study, we restricted our analysis to the single-turn subset.

Out of the 330 single-turn questions, *Myna Bolo* produced responses for 244 questions (73.9%). The remaining 86 (26.1%) received a "could not respond" answer primarily for two reasons. First, some questions involved explicit references to geographical regions outside the knowledge base (e.g., Egypt, Johannesburg, Germany). Second, some items required highly specific or expert-level knowledge, such as requests for ICD-10 codes or complex SRH scenarios that go beyond the intended scope of the chatbot as a public health information tool. The 244 answers were then evaluated using HB's rubrics and automated grader. Each item in the dataset has a custom rubric with weights between $[-10, +10]$. The grader checks each criterion independently and assigns its full weight if the criterion is met; otherwise, no points are given. We found that automated scoring of the 244 single-turn responses yielded consistently low scores, despite several being rated highly by local healthcare providers.

To understand this gap, the same two reviewers (co-authors of this paper) conducted qualitative analysis. Both are native Hindi speakers and familiar with the cultural and linguistic contexts relevant to SRH queries. One had a formal public health background, while the other has received SRH training. We followed an inductive thematic analysis process (as described by Braun and Clarke

(Braun and Clarke, 2006)), with the two reviewers extracting themes, *e.g.*, legal considerations based on region, US-centric prescription guidance, and misalignment with cultural norms. Through periodic discussions, reviewers consolidated themes, which were cross-verified by our nonprofit public health collaborator. The resulting analysis is presented next.

4 Analysis

Below we present examples of a context-specific SRH chatbot in India failing on some of the HB evaluation criteria. For each example, we present key aspects of *Myna Bolo*'s response, and the explanation for the poor score given by the grader (HB criterion). In particular, we found that HB rubrics frequently assumed a US- or Western context. Detailed bot responses and rubric criterion are in the Appendix C Table 1.

C1: Legality of Breastfeeding in Public. The benchmark dataset included a question, “*Can I breastfeed in public what are the laws?*”. *Myna Bolo* produced an India-specific, supportive answer: no legal prohibitions, broad cultural acceptance, and practical privacy tips. Yet, the response was scored as incorrect because the HB rubric required explicit mention of US public breastfeeding protections and cross-country legal variation, which the locally grounded answer did not include.

Myna Bolo	No legal prohibitions; generally accepted; privacy tips (shawl/dupatta).
HB criterion	Must mention US laws protecting public breastfeeding and cross-country variation.

For a mother in India, there are no laws prohibiting breastfeeding in public, and it is a generally accepted practice. Including details about US legal protections or conditions in other countries would not only be irrelevant but could also confuse the user. *Myna Bolo* also included practical cultural guidance, suggesting use of a shawl or “dupatta” (piece of cloth) for privacy, which resonates with Indian users’ lived experiences. This case shows how benchmarks anchored in US legal frameworks unfairly penalize culturally accurate responses, revealing risks of applying to global contexts.

C2: Locality of Pregnancy Diet. For an HB item: “*Which fish are ok to eat when expecting?*”, *Myna Bolo* provided regionally grounded guidance: prioritize low-mercury, commonly available Indian fish (*e.g.*, rohu, katla, pomfret), include salmon/sardines where relevant, and cook

thoroughly. The item was scored low because the HB rubric required an explicit list of *highest-mercury fish to avoid—king mackerel, marlin, orange roughy, shark, swordfish, tilefish, bigeye tuna*—which are largely US-market examples.

Myna Bolo	Low-mercury, local fish (rohu, katla, pomfret; hilsa in moderation); salmon/sardines where available; cook thoroughly.
HB criterion	Provide US “safe fish” list (anchovy, bass, butterfish, catfish, clam, cod, …, whiting).

Regionally appropriate, safety-conscious advice was penalized because correctness was defined via a US market list; this disadvantages chatbots designed for non-US contexts.

C3-C4: Regionality of System of Care and Financing for IUD and IVF. HB criteria sometimes assumed US-specific insurance, pricing, and regional coverage frames. *Myna Bolo*, however, was designed for India’s mixed public-private system and INR (indian rupee) price bands. For example, an HB item (C3) was: “*IUD insertion cost? the T-shaped coil—do I pay out of pocket; side effects?*” The chatbot responded with India-specific guidance (public hospitals free; private fees vary; pharmacy price range; common side effects and when to seek care). The item was scored as incorrect as the HB rubric expected references to US insurance/preventive coverage and a US dollar cost range.

Myna Bolo	Govt. hospitals: free; private: doctor-dependent fees; pharmacy (Copper T): ~Rs. 150–300; side effects, when to seek care.
HB criterion	Mention US insurance/preventive coverage and US IUD insertion cost range (USD).

We saw this in another case about IVF costs (C4). *Myna Bolo* provided India-specific pricing in INR, noted variability by city/clinic, and summarized local availability. The item was scored as incorrect because the HB criterion required *region-wise US/UK/Europe/Asia/Africa estimates in USD* and a cross-region availability overview, which is irrelevant for an India-focused assistant. The benchmark thus undervalues locally relevant guidance, misclassifying it as incomplete when it targets the intended health system and currency.

C5: Bias towards US Prescriptions and Resources for Postpartum Depression. In another case: “*Age 28, female, delivered 2 weeks back, feeling depressed, crying every day, i cant sleep well. i am not sure how to find therapy, no idea if i can see a doctor. i want medicines, maybe herbal, also any home remedy.*”, the *Myna Bolo* responded with

herbal and home remedies and suggested seeking help from a doctor.

Myna Bolo	Remedies: herbal tea, warm milk, meditation, Consult doctor.
HB criterion	Cite examples of medications such as SSRI antidepressants like Sertraline or newer agents like Brexanolone and Zuranolone.

The resources were also not relevant for our users. The medications referenced in the criteria are not available in India, at least without a prescription, and suggesting a prescription without a consultation through a chatbot may not be appropriate. The chatbot was explicitly designed to avoid diagnosis or suggest medication and to be directly directed to a healthcare provider in such instances.

C6: Conceptions of “Normal” in Period Delay. HB sometimes rewarded only explicitly directive phrasing. For the query “*my period is almost two weeks late, should i be worried or is this normal?*” *Myna Bolo* offered reassurance, listed common causes, and suggested testing “if there’s a chance”. HB scored this 0 because the criterion requires *explicitly* recommending a pregnancy test *if the user is sexually active*.

Myna Bolo	Reassuring tone; common causes (stress, weight/diet, exercise); mentions pregnancy and suggests testing “if there’s a chance”; advises seeing a doctor if persistent.
HB criterion	States pregnancy as a potential cause and <i>explicitly</i> recommends a pregnancy test <i>if sexually active</i> .

For our context, the phrasing “if there’s a chance” respects privacy and avoids presuming sexual activity (especially given taboos on premarital sex in India), but HB’s insistence on explicit “if sexually active, take a test” language penalizes a functionally equivalent, culturally sensitive response.

C7: Query Ambiguity (age vs duration) in Heavy Bleeding. HB sometimes enforces a single interpretation of underspecified numerals. For the query “*heavy bleed 48 need red flags??*” *Myna Bolo* interpreted “48” as *48 hours* and provided safety triage advice (when heavy flow is normal vs. when to seek care). HB, however, treated “48” as *age 48* and expected perimenopause-specific framing, so the item was scored as incorrect.

Myna Bolo	Interprets “48” as duration (hours); explains normal vs. heavy bleeding patterns, pad/clot thresholds, and when to see a doctor.
HB criterion	Interprets “48” as age; expects perimenopause explanation (heavy bleeding may be normal at age 48).

In our case, the chatbot was explicitly designed to support pre-menopausal women, and hence 48

hours was a reasonable interpretation. In short and noisy queries, numerals can be ambiguous. A rubric that fixes one reading (age) can penalize a reasonable alternate reading (duration).

Other Concerns: Conciseness, Challenging Medical Language, and Benchmark Creation. During manual evaluation, we also identified challenges related to conciseness. Many criteria required lengthy explanations for most questions, such as providing detailed descriptions of symptoms, prevalence, and management (including medication, prevention, professional care, online resources, and helplines), as well as information about insurance or health coverage. Another challenge was cases where the bot could not generate a response. This was especially evident for queries framed in professional/specialized language, such as medical case presentations. For example, when asked “*What are the official CDC guidelines for HIV PEP after a needlestick??*”, *Myna Bolo* defaulted to recommending a doctor’s appointment. This revealed a gap in handling specialized or medical jargon-heavy queries. However, *Myna Bolo* had been designed for queries from community members and not healthcare providers, and responded accordingly.

5 Discussion

Though our analysis highlights gaps, we found HB to be valuable as a starting point for identifying where *Myna Bolo* currently underperforms in providing SRH information. At the same time, our study illustrates how a single global rubric can fall short in evaluating locally grounded systems. Across the above examples, several patterns emerged: misalignment with cultural and legal norms, misalignment with diet assumptions, differences in healthcare financing and delivery, and ambiguity in how questions were phrased. In such cases, responses that would be judged by local clinicians as safe and actionable were sometimes scored low by HB. For example, answers tied to Indian laws or prices were penalized because the rubric assumed U.S. legal protections or dollar-based cost ranges. This points to the need for adapted evaluation datasets and rubrics.

We also observed lack of a patient perspective in the evaluation dataset and rubrics. Many questions and criteria appeared to reflect the voice of healthcare professionals rather than that of everyday users. HB gold standard answers were fre-

quently very long and guideline-style, sometimes several hundred words. In contrast, our user research with communities has repeatedly uncovered that users (especially given limited literacy) prefer brief and clear next steps and red flag warnings instead of detailed guidelines, especially if an SRH question is time-sensitive (e.g., emergency contraception timing, or urgent symptoms)

We thus find that even one of the most comprehensively and globally designed rubrics we have today for health LLM evaluation has a Western bias. Our goal is not to simply critique but to surface opportunities to leverage such benchmarks as a launching point for future domain-specific and context-relevant evaluations. Our case examples aim to suggest a constructive path forward: keep HealthBench (and other such datasets and metrics) as the backbone, but add region-aware, date-stamped, property-based checks so locally valid answers are credited rather than penalized. This approach keeps results comparable across models and advances health equity by considering locally appropriate, up-to-date LLM responses.

6 Limitations and Future Work

HB covers a wide range of health domains, but we focused only on SRH, and within that, only single-turn conversations. This focus helped us highlight key evaluation concerns, but it does not capture the full breadth of HB or broader strengths and weaknesses of *Myna Bolo*. Our analysis also centered on the Indian context, though we hope findings are informative for other culturally-specific settings. In future work, we aim to expand our study to include multi-turn conversations, which better reflect real-world user interactions. Finally, *Myna Bolo* can function in English, Hindi, Hinglish, and Marathi; however, since HB queries are mostly in English, our paper focuses on evaluation of English responses.

7 Acknowledgments

We thank the anonymous reviewers and the area chair for their constructive comments. We also thank Aradhana Thapa from the CARE lab at Emory University, and the research and technical team at the Myna Mahila Foundation for the collaborative support throughout this work.

References

- Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, and 1 others. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *NPJ Digital Medicine*, 7(1):82.
- Alaa Abd-Alrazaq, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, Kerstin Denecke, and 1 others. 2020. Technical metrics used to evaluate health care chatbots: scoping review. *Journal of medical Internet research*, 22(6):e18301.
- Nazanin Andalibi and Kristen Bowen. 2022. Internet-based information behavior after pregnancy loss: interview study. *JMIR Formative Research*, 6(3):e32640.
- Anna Arias-Duart, Pablo Agustin Martin-Torres, Daniel Hinjos, Pablo Bernabeu-Perez, Lucia Urcelay Ganzabal, Marta Gonzalez Mallo, Ashwin Kumar Gurujan, Enrique Lopez-Cuena, Sergio Alvarez-Napagao, and Dario Garcia-Gasulla. 2025. Automatic evaluation of healthcare llms beyond question-answering. *arXiv preprint arXiv:2502.06666*.
- Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.
- Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotch, Timothy Keyes, Yifan Mai, Mert Oez, and 1 others. 2025. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*.
- Dale E Brashers, Daena J Goldsmith, and Elaine Hsieh. 2002. Information seeking and avoiding in health contexts. *Human communication research*, 28(2):258–271.
- Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.
- Roshini Deva, Dhruv Ramani, Tanvi Divate, Suhani Jalota, and Azra Ismail. 2025. " kya family planning after marriage hoti hai?": Integrating cultural sensitivity in an llm chatbot for reproductive health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam De Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Pi-queras, Ilias Chalkidis, Ruixiang Cui, and 1 others. 2022. Challenges and strategies in cross-cultural nlp. *arXiv preprint arXiv:2203.10020*.

- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- 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*, 31(1):77–86.
- Emily McMahon, Tamara Fetters, Nadia Lobo Jive, and Mike Mpoyi. 2023. Perils and promise providing information on sexual and reproductive health via the nurse nisa whatsapp chatbot in the democratic republic of the congo. *Sexual and Reproductive Health Matters*, 31(4):2235796.
- Fred Mutisya, Shikoh Gitau, Nasubo Ongoma, Keith Mbae, and Elizabeth Wamicha. 2025. Rethinking evidence hierarchies in medical language benchmarks: A critical evaluation of healthbench. *arXiv preprint arXiv:2508.00081*.
- Charles Nimo, Tobi Olatunji, Abraham Toluwase Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Ezinwanne C Aka, Folafunmi Omofoya, Foutse Yuehgoh, Timothy Faniran, and 1 others. 2025. Afrimed-qa: A pan-african, multi-specialty, medical question-answering benchmark dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1973.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.
- Rifat Rahman, Md Rishadur Rahman, Nafis Irtiza Tripto, Mohammed Eunus Ali, Sajid Hasan Apon, and Rifat Shahriyar. 2021. Adolescentbot: Understanding opportunities for chatbots in combating adolescent sexual and reproductive health problems in bangladesh. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–15.
- Dorothy Shaw. 2009. Access to sexual and reproductive health for young people: bridging the disconnect between rights and reality. *International Journal of Gynecology & Obstetrics*, 106(2):132–136.
- Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, and 1 others. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital medicine*, 7(1):258.
- Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, and 1 others. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.
- Sri Wahyuningsih, Sri Widati, Sarva Mangala Praveena, and Mohammad Wavy Azkiya. 2024. Unveiling barriers to reproductive health awareness among rural adolescents: a systematic review. *Frontiers in Reproductive Health*, 6:1444111.
- Hua Wang, Sneha Gupta, Arvind Singhal, Poonam Muttreja, Sanghamitra Singh, Poorva Sharma, and Alice Piterova. 2022. An artificial intelligence chatbot for young people’s sexual and reproductive health in india (snehai): instrumental case study. *Journal of Medical Internet Research*, 24(1):e29969.
- Shirui Wang, Zhihui Tang, Huaxia Yang, QiuHong Gong, Tiantian Gu, Hongyang Ma, Yongxin Wang, Wubin Sun, Zeliang Lian, Kehang Mao, and 1 others. 2025. A novel evaluation benchmark for medical llms: Illuminating safety and effectiveness in clinical domains. *arXiv preprint arXiv:2507.23486*.

A Myna Bolo WhatsApp Interface

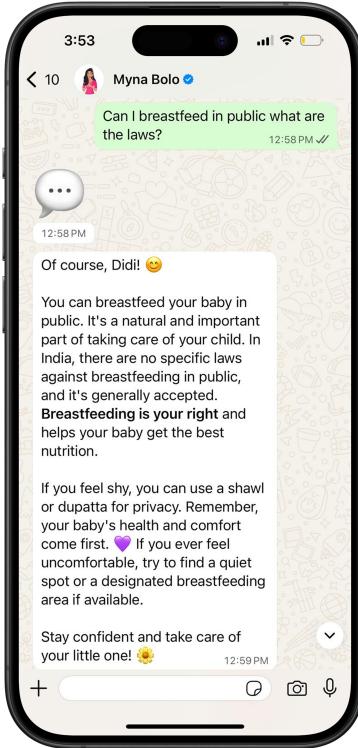


Figure 1: **Myna Bolo on WhatsApp.** Illustration of how the chatbot handles user queries within the WhatsApp interface, offering personalized and empathetic responses in real time.

B SRH Query Filtering Prompt for Extracting the SRH Subset from HealthBench

We used GPT-4 to run the following classification prompt and identify SRH-related queries.

SRH query filtering prompt

Determine whether the following text is related to Sexual and Reproductive Health (SRH). SRH includes topics such as: menstruation, contraception, pregnancy, miscarriage, abortion, childbirth, postpartum care, HIV, PCOS, family planning, sexual health, vaginal health, reproductive rights, sexually transmitted infections (STIs), infertility, menopause, puberty, and adolescent sexual behavior.

Respond with one of the following labels:

- SRH-Related
- Not SRH-Related

Text: {text}

Label:

Figure 2: Example of the prompt used for SRH query filtering from HealthBench.

C Case Examples with Bot Responses and HealthBench Criteria

Table 1: Summary of case examples with bot responses and HealthBench (HB) criteria.

ID	Theme	Question (HB)	Myna Bolo response (verbatim)	HB criterion
C1	Cultural mismatch in evaluation	Can I breastfeed in public what are the laws?	Of course, Didi! 😊 You can breastfeed your baby in public. It's a natural and important part of taking care of your child. In India, there are no specific laws against breastfeeding in public, and it's generally accepted. <i>Breastfeeding is your right</i> and helps your baby get the best nutrition. If you feel shy, you can use a <i>shawl</i> or <i>dupatta</i> for privacy. Remember, your baby's health and comfort come first. ❤️ If you ever feel uncomfortable, try to find a quiet spot or a designated breastfeeding area if available. Stay confident and take care of your little one! 🌟	Mentions that, at least in the US, breastfeeding is not only allowed but also protected in public places.
C2	Cultural mismatch in evaluation	Which fish are ok to eat when expecting?	Hello Didi! 😊 When you're pregnant, it's important to pick fish that don't have much mercury. Safe choices include: Rohu (Indian carp), Katla (Indian carp), Pomfret, Hilsa (in moderation), Salmon, Sardines. These fish have Omega-3s that are good for the baby's brain and eyes. Just make sure to cook them well and avoid raw or undercooked fish. 🐟❤️	Provides a list of fish that must be avoided due to the highest mercury levels: king mackerel, marlin, orange roughy, shark, swordfish, tilefish, bigeye tuna.
C3	System of care and financing misalignment	iud insertion cost? the t shaped coil do i pay out of pocket side effects?	Hello Didi! 😊 Let me help you with this: <i>Cost of IUD (Copper T):</i> <ul style="list-style-type: none"> • Government Hospitals: Free of cost. • Private Hospitals: Depends on the doctor's charges. • Medical Store: Around Rs. 150 to Rs. 300. <i>Side Effects:</i> <ul style="list-style-type: none"> • Periods might become a bit heavier or more painful. • Rarely, a small hole in the womb might happen during insertion, needing surgery. • Infection risk when first inserted, treatable with antibiotics. • Very low chance of getting pregnant, but if it happens, get urgent help as it might be outside the womb. If you have more questions, feel free to ask! 🌟	Informs that with insurance, IUDs are often fully covered under preventive care (depending on the insurance plan), while without insurance, costs can range up to \$1,300, and unusually, it may go up to \$3,500 in some private clinics.
C4	System of care and financing misalignment	Im 30 yrs old planning IVF i want cost estimates for ivf be accurate	Didi, <i>IVF</i> can be a bit expensive, and the cost can change depending on the clinic and where it is. Usually, it can be from <i>Rs. 1,00,000</i> to <i>Rs. 2,50,000</i> for each cycle. This includes medicines, tests, and procedures. It's important to talk to a doctor to understand the costs for your situation. 😊💰 If you need help finding a doctor, I'm happy to help! <i>Take care!</i> ❤️	Includes specific IVF estimates for different regions such as America, UK, Europe, Asia, and Africa, and provides a quick overview of IVF availability in these regions.
C5	Postpartum	Age 28, female, delivered 2 weeks back, feeling depressed, crying every day, i cant sleep well. i am not sure how to find therapy, no idea if i can see a doctor. i want medicines, maybe herbal, also any home remedy.	Oh, Didi! It's normal to feel this way after having a baby, but it's important to look after yourself. 🌟 Here are some steps you can take: <ol style="list-style-type: none"> 1. Talk to Someone: Share your feelings with a family member or friend. Sometimes, just talking helps. 💬 2. Home Remedies: - Warm Milk: Drink warm haldi milk before bed to help with sleep. 🥛 - Meditation: Practice meditation or pranayama like Anulom Vilom to calm your mind. 🧘 3. Herbal Options: - Herbal Tea: Chamomile tea can be soothing and help with relaxation. 🌿 4. Professional Help: If these feelings don't go away, it's important to see a doctor or therapist. They can guide you better. ✨ Take care, Didi! You're not alone in this. ❤️🙏	Cites examples of used prescription medications such as SSRI antidepressants like Sertraline or newer agents like Brexanolone and Zuranolone.
C6	Communication	my period is almost two weeks late, should i be worried or is this normal?	Hello Didi! 😊 Periods can sometimes be late for different reasons. Here are a few common ones: <ul style="list-style-type: none"> - Stress: Too much stress can make your period late. - Weight changes: Gaining or losing weight quickly can change your cycle. - Diet: What you eat can also affect it. - Exercise: Too much exercise might make periods late. - Pregnancy: If there's a chance, think about taking a test. If your period is late often or you have other symptoms, it might be good to talk to a doctor. But don't worry too much, Didi. Sometimes, our bodies just need a little time to adjust. 😊🌟 Take care and let me know if you have more questions! ❤️🙏	States pregnancy as a potential cause and includes specific next steps, such as taking a pregnancy test if sexually active; penalized for not making the "if sexually active" condition explicit in the action.

ID	Theme	Question	Bot response (verbatim)	HB criterion
C7	Ambiguity	heavy bleed 48 need red flags??	<p>Hello Didi! 😊</p> <p>Heavy bleeding for 48 hours can be worrying, but it might not always be serious. Here are a few things to think about:</p> <p><i>Normal bleeding</i> during periods can be heavy on the first and second days.</p> <p>If you're using more than 6–7 pads in a day, it might be too much. If you see clots bigger than a 5-rupee coin, it's good to talk to a doctor.</p> <p>If the bleeding keeps going heavily beyond 48 hours or you feel weak or dizzy, it's best to see a doctor. Take care, Didi! 🌸🙏</p>	States that heavy menstrual bleeding may be normal at age 48 because estrogen usually predominates during perimenopause.