
Evaluating the Early Deployment of a GenAI Co-Pilot for Frontline Maternal and Newborn Health Communication: A Mixed-Methods Case Study from India

Zezhen Wu¹ Poornima Sharma² Shirley Du Yan³ Jamie Sewan Johnston⁴

Victoria C. Ward⁴ Gary L. Darmstadt⁴ Aman Dalmia¹

¹The Agency Fund ²Noora Health India Private Limited ³Noora Health ⁴Stanford University
{michael,aman}@agency.fund {poornima,shirley}@noorahealth.org {jamiejs,vward,gdarmsta}@stanford.edu

Health systems in low- and middle-income countries are beginning to use generative AI (GenAI) to help frontline workers manage high volumes of incoming messages [7]. Because these tools often reach patients through human workers, evaluations should examine how GenAI affects workers' judgment, communication practices, and agency, and the services patients receive.

A nonprofit health organization and its partners in India operate a two-way WhatsApp service that extends patient and caregiver health education beyond the facility; trial evidence shows such education improves postnatal knowledge and practices [4]. It receives about 240,000 queries per month (18,000 medical, 750 emergency) from 80,000 users across 15,000 facilities. In November 2025, the organization introduced a GenAI co-pilot to a subset of its Medical Support Executives (MSEs), who answer messages in Hindi, Telugu, Punjabi, Kannada, Odia, and Marathi. The co-pilot drafts a language-matched reply to each query from documented patient history and context; the MSE reviews, edits, and sends it. The tool is not patient-facing; its effects run through how MSEs handle its suggestions.

The study asks two questions. **RQ1:** How do MSEs with co-pilot access experience the tool, including changes in attitudes, trust, agency, and practice, and what do MSEs without access expect from it? **RQ2:** Are conversations handled by co-pilot MSEs measurably different in interactional dynamics, informational content, and rubric-rated quality, and do they match MSEs' reported experiences?

The design is concurrent mixed methods [2]. For RQ1, we conduct 18 semi-structured interviews, nine per group, analyzed with reflexive thematic analysis [1], paired with interaction logs of each query, AI draft, and sent message; edits to drafts trace where the system meets or misses patients' needs. For RQ2, we build measures that apply to any conversation, with or without co-pilot use. Researchers and clinicians code a stratified sample of about 200 conversations in multiple languages, combining several evaluation rubrics (e.g., clarity, empathy, medical safety, accuracy, next-step guidance) [6] with open coding for patterns the rubrics miss, such as culturally specific markers. We pay special attention to the annotation process because prior efforts report inter-rater agreement below 50%; documenting it carefully yields methodological insight for AI evaluation in conversational health education, where data are often fragmented and heterogeneous. We then validate an LLM-judge pipeline against human consensus ratings co-developed by clinical experts and medical doctors and apply it to the full corpus [3, 8]; clinicians manually review conversations flagged for safety risks to check alignment. We also compute interactional metrics (latency, length, turn count, follow-up) and informational metrics (reading complexity, clinical terminology) from logs. Before launch, 18 India-based MSEs were randomized 1:1 to treatment or control, stratified by language. Once conversations are scored, we estimate whether those handled by co-pilot MSEs differ from control, using multilevel models with conversation-level outcomes, MSE random intercepts, and small-sample corrections for the 18 clusters [5]. The protocol was approved by the ACE Independent Ethics Committee, Bangalore.

For AI evaluation in the Global South, our case study makes two contributions. First, it offers a practical example for evaluating human-in-the-loop GenAI in multilingual settings where ground-truth labels are scarce and outcomes depend on frontline workers' judgments. Second, it also shows how evaluators can triangulate workers' reported experience against behavioral traces in service data, testing whether attitudes toward GenAI match changes in communication.

References

- [1] Virginia Braun and Victoria Clarke. *Thematic Analysis: A Practical Guide*. SAGE, London, 2022.
- [2] John W. Creswell and Vicki L. Plano Clark. *Designing and Conducting Mixed Methods Research*. SAGE, Thousand Oaks, CA, 3rd edition, 2018.
- [3] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.
- [4] Jamie Sewan Johnston, Pooja Suri, Shirley Yan, Adithi Chandrasekar, Saumya Singla, Victoria C. Ward, and Seema Murthy. A mobile messaging service for families on postnatal knowledge and practices: a cluster randomized trial, India. *Bulletin of the World Health Organization*, 103(4): 255–265, 2025.
- [5] Michael G. Kenward and James H. Roger. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53(3):983–997, 1997.
- [6] Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V. Stolyar, Katelyn Polanska, Karleigh R. McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, Piyush Mathur, Giovanni E. Cacciamani, Cong Sun, Yifan Peng, and Yanshan Wang. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7:258, 2024.
- [7] World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*. World Health Organization, Geneva, 2024.
- [8] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 2024.