1. Overview

1.1 Organization Name

ARMMAN

1.2 Domain

Maternal health – Frontline health worker support

1.3 Use case / problem being addressed

On-demand guidance for Auxiliary Nurse Midwives (ANMs) to interpret high-risk pregnancy conditions, recall clinical protocols, and support case management during antenatal care and follow-up

1.4 Target population

- User: Auxiliary Nurse Midwives
- Geography: India (Uttar Pradesh and Telangana)
- Language: Hindi and Telugu. Note that users primarily interact in their state language with English mixed in. English is primarily used for numbers, medical terminology, or commonly adopted health-related terms.

1.5 Deployment stage

Production

2. Dataset

2.1 Data source and collection

The data consists of questions from real users, i.e., ANMs in Uttar Pradesh and Telangana, from 1 January 2026 to 31 March 2026. It also has the answer rubric for each question.

2.2 Data composition

- Size: 1010 questions
- Modality: Text only

- Language(s) and dialect(s): Hindi and Telugu
- Distribution: 805 Hindi, 205 Telugu questions
- Sampling logic: We sampled 1010 questions from six common risk topics (General antenatal care, general high-risk pregnancy, anemia, pregnancy-induced hypertension, gestational diabetes, and fever in pregnancy. Only questions of length of at least 10 characters were included.

2.3 Limitations

- Only text questions included, which may have excluded questions from less digitally-literate ANMs who are more comfortable using voice notes to ask questions
- Only questions from six topics included: general antenatal care, general high-risk pregnancy, anemia, pregnancy-induced hypertension, gestational diabetes, and fever in pregnancy. Note that these topics account for over 80% of real user queries

2.4 PII

- There is no PII included (ANM name or phone number)

2.5 Unit of evaluation

Query (text only)

2.6 Data structure

Term	Description	Example
ID	int or string	msg_001
user_query	user's query asked to bot	Mahila Ka Sath gram blood Hai To Kya Karna chahie
query_category	category of query out of various high risk categories	ANEMIA
response	bot response to user's query	- अगर महिला का रक्त में Haemoglobin (Hb) स्तर 7 ग्राम/डेसीलीटर है, तो यह मध्यम (moderate) रक्ताल्पता (anemia) है। - *ऐसी स्थिति में Auxiliary Nurse and Midwife (ANM) को महिला को तुरंत Medical Officer (MO) या उच्च केंद्र पर रेफर करना चाहिए।*

		- साथ ही, महिला को आयरन और फोलिक एसिड (Iron and Folic Acid - IFA) टैबलेट्स नियमित रूप से लेने, संतुलित आहार लेने, और Albendazole 400 mg की एक खुराक (पहली तिमाही के बाद) देने की सलाह दी जाती है। - *यदि एक महीने बाद Haemoglobin (Hb) में 1 ग्राम/डेसीलीटर की वृद्धि नहीं होती है, तो फिर से Medical Officer (MO) को रेफर करें।* - रक्ताल्पता के लक्षण जैसे थकान, सांस फूलना, चक्कर आना आदि पर भी ध्यान दें।
rubrics	The benchmarks/rules, on which bot response will be graded. its structure is JSON.	{ clarity: "The response should be simple, direct, and action-focused, using clear frontline language. It should avoid long protocol details and make the referral decision unmistakable. Minor grammar issues are acceptable if the meaning remains clear.", completeness: "ESSENTIAL: State that Hb 7 g/dL in pregnancy is moderate anaemia and needs immediate referral to MO/higher centre. Mention u…ce, diet advice, and repeat Hb follow-up may be discussed after evaluation, but these are not substitutes for referral now.", factual_correctness: "A satisfactory answer must interpret “7 gram blood” as Hb 7 g/dL in a pregnant woman, which is moderate anaemia. It must cor…erral to a Medical Officer or higher centre. It should not say this can be managed only with routine ANM treatment at home." }
factual_correctness	one of the benchmark of above rubric	A satisfactory answer must interpret “7 gram blood” as Hb 7 g/dL in a pregnant woman, which is moderate anaemia. It must cor…erral to a Medical Officer or higher centre. It should not say this can be managed only with routine ANM treatment at home.
completeness	one of the benchmark of above rubric	ESSENTIAL: State that Hb 7 g/dL in pregnancy is moderate anaemia and needs immediate referral to MO/higher centre. Mention u…ce, diet advice, and repeat Hb follow-up may be discussed after evaluation, but these are not substitutes for referral now.
clarity	one of the benchmark of above rubric	The response should be simple, direct, and action-focused, using clear frontline language. It should avoid long protocol details and make the referral decision unmistakable. Minor grammar issues are acceptable if the meaning remains clear.
lang	language of user query or bot response	hindi

3. Evaluation design

3.1 What is being measured

- Our evaluation rubric had 3 dimensions, with Yes/No answers
    - Factual correctness: Is the response factually and medically correct, with no incorrect or misleading information?
        - Yes: All medical or factual information in the response is correct and aligned with accepted clinical guidance.
        - No: The response includes any incorrect, misleading, or medically inaccurate information (e.g., incorrect advice, wrong cutoff values, incorrect HRP classification, etc)
    - Completeness: Does the response provide all the important information needed to answer the question appropriately?
        - Yes: The response includes all the key information needed to answer the question appropriately, including any important next steps or warnings.
        - No: Important information related to the question is missing, such as key advice / guidance or referral cues that the user would reasonably need.
    - Clarity: Is the response clear, understandable, and appropriately phrased for the intended user (ANM)?
        - Yes: The response is easy to understand, clearly worded, logically structured, and appropriate for the intended user
        - No: The response is confusing, poorly phrased, hard to follow, overly technical/formal, or not well structured.

3.2 Rationale

- We started with a larger set of evaluation dimensions: relevance, factual correctness, harm safety, completeness, absence of irrelevant content, clarity, and overall satisfactoriness.
- However, there was significant overlap between the dimensions. The three dimensions finally chosen were the most comprehensive, as well as easiest to capture as rubrics

4. Labels and annotation

4.1 Overall labelling process

- 100 Q&A pairs (50 in Hindi, 50 in Telugu) were labeled by 2 annotators each along 6 dimensions: relevance, factual correctness, harm safety, completeness, absence of irrelevant content, clarity, and overall satisfactoriness.
    - 2 annotators were specialist doctors (gynaecologists), 1 was a doctor with public health experience, 1 was a nurse-trainer with public health experience
    - Each Q&A pair got 12 (6x2) labels, along with comments on cases where the answer got a negative label
- Guidelines provided to annotators: LINK
- Inter-annotator agreement (if measured) — methodology and results:
    - On the set of 50 Hindi questions, we had 100% inter-annotator agreement on factual correctness, 96% on completeness and 84% on clarity
    - On the set of 50 Telugu questions, we had 90% inter-annotator agreement on factual correctness, 76% on completeness and 80% on clarity
- This exercise was used to design an LLM judge to generate the rubrics, by providing examples of acceptable and unacceptable responses
- The LLM judge generated rubrics for 1000 questions (800 Hindi, 200 Telugu). 50 rows (questions + rubrics) for each language were validated by human annotators. The rubric generator prompt was refined based on this feedback
- We repeated this exercise after running the full benchmark pipeline on the actual responses shared with users. The LLM judge's scores and comments were used to further refine the rubric wherever the judge rated ARMMAN's ANM chatbot answers as unsatisfactory on any of the dimensions but human evaluators found the response satisfactory–in these cases, either the rubric was made more lenient, or the judge model was refined to be more lenient

4.2 LLM-as-judge

- Model used as judge: gpt-5.4
- Whether judge model differs from generation model: Yes judge model is more capable than generation model (gpt-4.1), due to cover subteles of rubric compliance in generated response.
- Validation against human labels — From the presented dataset, we selected 100 pairs (50 Hindi, 50 Telugu) for human review. Each pair was evaluated by two independent experts on three parameters: Factual Correctness, Completeness, and Clarity (Yes/No).

5. Benchmark

5.1 Cross-cutting dimensions

Cost: Not available

Latency: Not available

Accuracy:

factual_correctness_avg_score	completeness_avg_score	clarity_avg_score
73.83	42.08	86.63

- Safety and guardrail checks: redirect to medical person when out of scope or unsure

5.2 How the benchmark runs

- Inputs the benchmark takes: user_query, bot_response, rubrics
- Outputs the benchmark produces: aggregate score, individual query level score, reasoning of score.

5.3 Reproducibility

- Code availability and link: “https://github.com/armman-ngo/ANM_Assistant_Benchmarking/tree/final_submission”
- Dependencies and environment requirements: available in github repository
- Evaluation prompts and judge configurations (full values in the code): available in github repository
- Methodological repeatability — what someone needs to replicate this: available in github repository

5.4 Position in the Evaluation Ecosystem

- Related benchmarks (mention citations) - Not evaluated
- Relationship to existing frameworks (L1-L4, HealthBench, etc.) — adopted, adapted, or distinct (if any) - Not evaluated
- Gap this benchmark fills/how this differs - Not evaluated

6. Worked Example

Following data is produced by our process. It shows aggregate/average scores of factual_correctness, completeness, completeness as well as query level scores and reasoning of scoring. In the following only one output is shown at query level. Since we are using bot response as it was (not re generating) and we did not collect latency and cost earlier, so inference_latency_sec and inference_cost_usd are not available to us.

total_samples_evaluated	1010
average_inference_latency_sec	0
average_inference_cost_usd	0
aggregate_pass_rates	
factual_correctness	"73.8336713995943%"
completeness	"42.0792079207921%"
clarity	"86.6336633663366%"
itemized_results	
id	"msg_001"
user_query	"Mahila Ka Sath gram blood Hai To Kya Karna chahie"
prediction	"- अगर महिला का रक्त में Haemoglobin (Hb) स्तर 7 ग्राम/डेसीलीटर है, तो यह मध्यम (moderate) रक्ताल्पता (anemia) है।\n- *ऐसी स्थिति में Auxiliary Nurse and Midwife (ANM) को महिला को तुरंत Medical Officer (MO) या उच्च केंद्र पर रेफर करना चाहिए।*\n- साथ ही, महिला को आयरन और फोलिक एसिड (Iron and Folic Acid - IFA) टैबलेट्स नियमित रूप से लेने, संतुलित आहार लेने, और Albendazole 400 mg की एक खुराक (पहली तिमाही के बाद) देने की सलाह दी जाती है।\n- *यदि एक महीने बाद Haemoglobin (Hb) में 1 ग्राम/डेसीलीटर की वृद्धि नहीं होती है, तो फिर से Medical Officer (MO) को रेफर करें।*\n- रक्ताल्पता के लक्षण जैसे थकान, सांस फूलना, चक्कर आना आदि पर भी ध्यान दें।"
scores	
factual_correctness_score	1
completeness_score	0
clarity_score	1
factual_correctness_reasoning	"The response correctly interprets 7 gram blood as Hb 7 g/dL in a pregnant woman, identifies it as moderate anaemia, and clearly advises immediate referral to a Medical Officer or higher centre rather than home-only management."
completeness_reasoning	"It includes the essential point that Hb 7 g/dL in pregnancy is moderate anaemia needing immediate referral, but it does not mention the specific danger signs that make urgent referral especially important: breathlessness, palpitations, severe pallor,

	pulse >100, or RR ">24.
clarity_reasoning	"The answer is direct and understandable, with the referral decision stated clearly and prominently. Despite a few extra management details, the main action remains unmistakable."
query_category	""
lang	""
inference_latency_sec	0
inference_cost_usd	0

7. Results

7.1 Models evaluated

- List of models tested: gpt-4.1
- Versioning and date of testing: 29-June-2026

7.2 Headline results

- Performance per dimension per model: mentioned in section 5.1
- Notable findings or failure modes:
    - The lowest performance is on completeness, where the model seems to omit information defined as essential by the rubric
    - On factual correctness too, most errors seem to be errors of omission.
    - On clarity, failure modes include use of overly formal language or complex medical jargon which may not be suited for frontline health workers. This was observed across both languages

7.3 Interpretation

- What the results tell us about model fitness for this use case
    - While the model fell short on completeness, factual correctness scores were decent and clarity scores remained high
    - Through human validation, many of the gaps on factual correctness and completeness were observed to be errors of omission, which human evaluators did not assess as critical errors.
    - The benchmark appears to have a high false negative rate (i.e., clinically acceptable answers rated as unacceptable because of minor gaps and omissions) and a very low false positive rate

- Where models fail and why it matters
    - Our GPT4.1 implementation was found to be lacking on the completeness of the information, which could have happened because of an emphasis on conciseness and easy readability
    - This points to a trade-off between perceived usefulness by field workers (necessary for adoption) and clinical expectations for pregnancy-related questions

8. Public Release

- How much is being released publicly (sample size): None
- How will you license the public sample: N/A

9. Full release for TAF/Endless Health

- Full dataset as a zip file along with scripts to run inference and automated evaluations and clear instructions for reproducibility

10. Acknowledgements and Contributors

Team members: Awadhesh Srivastava, Amrita Mahale, Sid Parida, Jigar Doshi, Nithesh S

Annotators (anonymized if needed): Vaishali Suyal, Dr Sowjanya Velugunbuntla, Dr Sowjanya Veeradasu, Dr Deepika Asthana

Funding acknowledgment: We thank Agency Fund and Endless Health for their support.