
Auditable Emergency Triage for Maternal and Newborn Care in India

Aman Dalmia^{2*}, Shobhit Jagga^{1*}, Niharika Priyadarshini^{1*}, Anubhav Arora¹,
Amrita K Prasen¹, Nikhil Nalin¹, Santhosh SJ¹, Sreeram Nurani Ramasubramanian¹,
Muhammed Afeer K¹, Neelima Devadas¹

¹Noora Health

²The Agency Fund

Abstract

We run a WhatsApp-based service where nurses answer over 50,000 medical queries per month. Their most time-critical task is emergency triage: deciding which queries need immediate in-person attention. To support them, we built a system that uses a large language model (LLM) to classify whether a message is an emergency and provide a rationale for interpretability. Nurses can flag if the system missed an emergency or incorrectly labeled it as one, giving us a live measure of missed emergencies and false alarms. But the system was opaque: analyzing mistakes meant reading reasoning chains for each message, infeasible at our scale. Prompt changes meant re-running a full evaluation to prevent regressions, both costly and operationally challenging. Clinicians follow a decision tree to make this call, but it was never documented or passed to the model, which relied on a flat list of danger signs. To address the issues, we decomposed triaging into two steps: an LLM extracts canonical symptoms and patient context from the query against a clinician-authored vocabulary, and a deterministic rule engine captures the scenarios that indicate an emergency. We show that the new system raises recall from 0.565 to 0.810 and F1 from 0.606 to 0.702, with structured rules driving most of the accuracy gains while the decomposition provides auditability: clinical experts inspect each stage of the new system to see whether the query was mis-translated, symptoms were incorrectly extracted, patient context wrongly inferred, or the necessary rules were missing. They add new rules independently without causing regressions and avoid running costly evaluations. Since deployment, the new system has triaged 152,421 patient queries and flagged 28,535 (18.7%) as emergencies. The over-escalation rate was 17.8% without any increase in missed emergencies. Clinicians have also added 48 new rules since deployment, evidence of the faster correction loop we set out to build.

1 Introduction

Noora Health is a nonprofit organization supporting a caregiver-education program in district hospitals across nine Indian states. Families receive training in essential maternal and neonatal care during hospitalization and post-discharge guidance from nurses and doctors through a two-way WhatsApp service that receives over 50,000 medical queries monthly. Nurses answer routine questions and refer emergencies for immediate in-person care. But identifying them timely and reliably is hard.

LLMs are increasingly used in healthcare, including multilingual, low-resource settings where demand for clinical guidance exceeds capacity [Thirunavukarasu et al., 2023, Chen et al., 2025]. Our messages make triage harder: they are short, code-switched [Khullar et al., 2025, Agarwal et al., 2025], describe symptoms in colloquial terms rather than clinical vocabulary [Zeng and Tse, 2006], and often omit context, although the same symptom can be an emergency or not depending on it:

*Equal contribution.

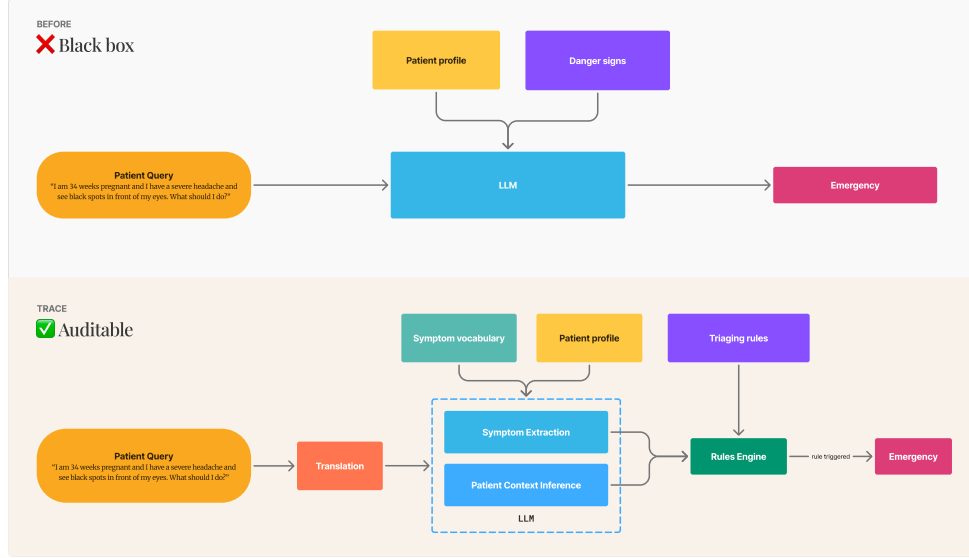


Figure 1: Evolution of the triage system. The initial version (top) uses an LLM to decide end-to-end whether a query indicates an emergency, with a list of danger signs as its knowledge base. TRACE (bottom) splits the task into LLM-based symptom extraction and patient context inference, followed by a deterministic rule engine, so every decision traces to the rule that drove it.

fever in a newborn is an emergency, while in an older child it usually is not. The stakes are high in India, which recorded an estimated 19,000 maternal deaths in 2023 and accounts for roughly 18% of neonatal deaths globally [World Health Organization et al., 2025, UN Inter-agency Group for Child Mortality Estimation, 2024], with delays in recognizing danger signs among the leading preventable causes.

We initially built our AI-assisted triage system by prompting an LLM, along with a list of common danger signs associated with an emergency, to classify the message and generate a reasoning chain for its decision (Figure 1). Nurses can see which messages have been flagged and they can override a wrong prediction, which gives us a continuous measure of over-escalations and missed emergencies in deployment. Since missed emergencies are more harmful, the system was optimized for higher recall. It caught real emergencies faster, but the number of over-escalations at our scale became a challenge.

The system was opaque. The only way to analyze mistakes was to read the reasoning generated by the LLM one message at a time, which is not practical at our scale. It also did not match how our clinicians work: they follow a decision tree that factors in patient context and symptom combinations, learned from training and experience rather than documented for a model to use. Our prompt carried only a flat list of danger signs, so any change risked degrading performance elsewhere and required a full evaluation before deployment, both costly and operationally challenging. These problems share a common root cause: the model was doing two jobs at once, interpreting the patient query and making the clinical decision.

2 System design

To address this, we built TRACE (**T**riage by **R**ule-based **A**pplication on **C**ontext-aware **E**xtractions). It separates the two jobs as a clinician would: an LLM extracts canonical symptoms from a fixed, clinician-curated vocabulary along with patient context such as trimester or infant age, and a deterministic rule engine, owned by the clinical team, decides whether to assign the emergency label. Symptom extraction is a narrower task than triage and is the only step that requires subjective interpretation of language. Once symptoms and context are extracted, the rules are fixed. Figure 1 compares the original end-to-end LLM system with TRACE.

The system triages one message at a time. We formalize the input as a tuple (m, c, P) . Here, m is the patient message. c is the care category under which the patient is enrolled with us, out of the

eight we support: antenatal care (ANC), postnatal care (PNC), special newborn care unit (SNCU), high-risk pregnant women (HRPW), general health and wellness (GHW), non-communicable disease (NCD), maternal and child health (MCH), and general medicine and surgery (GEMS). P is the profile information we maintain for every patient. For a pregnant patient, it may contain the expected delivery date and the gestational trimester $t \in \{1, 2, 3\}$ she is currently in. For a patient who has already delivered her baby, it stores the date of delivery and the infant’s age instead. Triage cannot be done from the message alone: the urgency of the same symptom varies with the patient context. The output is $y \in \{\text{emergency}, \text{non-emergency}\}$.

Queries span seven languages and are usually typed in Roman script rather than the native one, which degrades quality on this task [Khullar et al., 2025]. So, every query is first translated into English using Gemini-2.5-Flash. An LLM then extracts the canonical symptoms the message describes, along with any patient context it contains, and a short rationale. A patient may express the same symptom in many ways, often colloquially rather than clinically, so clinicians built a vocabulary mapping 745 patient phrasings to 184 canonical symptoms, following the practice of bridging everyday language to clinical concepts [Zeng and Tse, 2006]. A few examples are shown in Table 2 (Appendix). The rule engine then evaluates those symptoms and context together with c and P . The LLM never takes the final emergency (or not) decision. Clinicians prepared 231 rules. 141 are *Standalone*: the symptom on its own makes the message an emergency, whatever the patient context. The other 90 are *Combination* rules, where a symptom counts as an emergency only alongside a second symptom or a particular patient context, such as a given trimester or infant age. Abdominal pain alone, for example, is not necessarily an emergency, but abdominal pain with vaginal bleeding is flagged. Table 3 (Appendix) shows a few examples. This decomposition helps trace every mistake to the respective step: translation, symptom extraction, patient context inference, and rules.

3 Dataset and experiments

To evaluate our system, we built a dataset of 769 real patient queries with 251 emergencies, across all the languages we support: Hindi, English, Telugu, Kannada, Punjabi, Marathi, and Odia. The data distribution can be seen in Tables 4 and 5 (Appendix). It is divided into 342 validation queries (104 emergencies) and 427 unseen test queries (147 emergencies). Two in-house clinicians independently reviewed each query using the same information that was available to the system. They agreed on 83% of labels (Cohen’s $\kappa = 0.626$). Of the 130 queries where they disagreed, the senior clinician labeled 73 as non-emergency and the remaining 57 as emergency. Most of them were about whether a symptom such as pain, swelling, or reduced fetal movement was severe enough to be an emergency, not whether it was present. Disagreements were resolved by taking the senior clinician’s label as the ground truth.

We evaluate four architectures as a stepwise ablation to isolate the contribution of each design choice: **LLM-KB**, the baseline end-to-end LLM classifier using a flat list of danger signs; **LLM-Rules**, same as LLM-KB except the list of danger signs replaced by the structured rules; **TRACE-NoCtx**, the TRACE architecture without any patient context inference; and the complete **TRACE** architecture. Recall is the primary safety metric because a missed emergency is more harmful than a false alarm. F1 is the secondary metric for balancing recall with false alarms.

The same LLM, Gemini-2.5-Flash, and the same translated queries were used for every experiment. Since LLMs are stochastic, their outputs can vary on each run. To verify that our results are not driven by that variation, we ran the same configuration for each architecture three times on the validation set before running them on the test set. F1 ranged from 0.764 to 0.769 (mean 0.767, SD 0.002) and recall from 0.861 to 0.870 (mean 0.867, SD 0.005). The variation across runs is much smaller than the differences across architectures we observe. We also ran the same ablations with Gemini-3.5-Flash, Gemini-3.5-Flash-Lite, and Gemini-3.1-Pro on the validation set and found that the pattern holds. So, we report the results for only one LLM for the sake of brevity.

4 Results

Table 1 shows the comparison across the architectures on the test set.

Table 1: Comparison of the architectures on the test set ($N = 427$, 147 emergencies) with Gemini-2.5-Flash as the LLM. Brackets show 95% stratified-bootstrap confidence intervals. Structured rules drive most of the accuracy gains, and TRACE matches LLM-Rules within overlapping intervals while making each decision auditable (Section 5).

Approach	P	R	F1
LLM-KB	.654 [.586, .724]	.565 [.483, .646]	.606 [.539, .669]
LLM-Rules	.586 [.541, .633]	.837 [.776, .891]	.689 [.645, .732]
TRACE-NoCtx	.626 [.577, .679]	.810 [.741, .871]	.706 [.659, .752]
TRACE	.620 [.571, .672]	.810 [.741, .871]	.702 [.655, .748]

Formalizing rules drives the largest performance gain. Replacing the flat list of danger signs with structured rules raises recall from 0.565 to 0.837 and F1 from 0.606 to 0.689. The rules form the decision tree that our clinicians follow. Without them, the model relies on its own reasoning, shaped by its training data rather than our protocols.

Comparable performance with an auditable decision path. TRACE reaches a recall of 0.810 [0.741, 0.871] and F1 0.702 [0.655, 0.748], against 0.837 [0.776, 0.891] and 0.689 [0.645, 0.732] for LLM-Rules. Since the confidence intervals overlap, we cannot say one architecture is more accurate on this dataset, with TRACE providing the additional benefit of audibility as described in Section 5.

Context inference does not improve aggregate performance. TRACE-NoCtx and TRACE both reach a recall of 0.810, and their F1 intervals almost entirely overlap (0.706 [0.659, 0.752] against 0.702 [0.655, 0.748]). However, it remains necessary in practice. For example, a patient enrolled under GHW wrote that she was five months pregnant and had severe lower abdominal pain. TRACE-NoCtx kept the stored GHW category, found no applicable rule, and labeled the query non-emergency. TRACE inferred the true care category as ANC with the mother being in the second trimester, fired the matching rule and correctly classified the query as an emergency.

5 Auditability in deployment

Every error is traceable to a step. The decomposition maps a mistake to a mistranslation, a symptom either not extracted or missing from the vocabulary altogether, the wrong patient context inferred, or a rule that was wrong or missing. **Each component has an owner.** Clinicians handle missing rules and symptoms; extraction errors are analyzed together with the ML team. **Errors are resolved faster.** Since deployment, TRACE has triaged 152,421 patient queries and flagged 28,535 (18.7%) as emergencies. Nurses marked 17.8% as over-escalation and did not report more missed emergencies than before. As rule changes do not touch the LLM prompt anymore, clinicians can easily add new rules without running any expensive evaluations. Since deployment, they have added 48 rules on their own.

6 Limitations

Given the size of the test set, the metrics carry meaningful variance, making it hard to distinguish the finer differences between the architectures. The dataset is labeled following our internal protocols for the patients we serve, so the results may not generalize to other settings. An emergency whose symptoms are missing from the vocabulary map, or one that reflects a clinical scenario not yet covered by the rules, or requires a specific patient context missing from the query, will not be caught. All experiments were done on models from the Gemini family only.

7 Ethical statement

Our system is a nurse decision support rather than a patient-facing or autonomous system: nurses and doctors review every escalation and hold final authority because a missed escalation can cause serious harm, while over-escalation raises patient anxiety and load on an already overburdened public health system.

Acknowledgments and Disclosure of Funding

We thank the larger Noora Health team, participating nurses and clinicians, partner hospitals, government stakeholders, and the patients and families whose participation made this work possible.

References

- Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, Tao Sheng, Sujith Ravi, and Dan Roth. Aligning LLMs for multilingual consistency in enterprise applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 117–137. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-industry.9. URL <https://aclanthology.org/2025.emnlp-industry.9/>.
- Haichao Chen, Dian Zeng, Yiming Qin, Zeyue Fan, Faye Ng Yu Ci, David C. Klonoff, John S. Ji, Shuyang Zhang, Kwesi Nyan Amissah-Arthur, Michelle María Jiménez de Tavárez, Saleha Masood, Phuoc Van Le, Pearse A. Keane, Bin Sheng, Tien Yin Wong, and Yih Chung Tham. Large language models and global health equity: A roadmap for equitable adoption in LMICs. *The Lancet Regional Health – Western Pacific*, 63:101707, 2025. doi: 10.1016/j.lanwpc.2025.101707.
- Manurag Khullar, Utkarsh Desai, Poorva Malviya, Aman Dalmia, and Zheyuan Ryan Shi. Script gap: Evaluating LLM triage on Indian languages in native vs romanized scripts in a real world setting, 2025. URL <https://arxiv.org/abs/2512.10780>.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8): 1930–1940, 2023. doi: 10.1038/s41591-023-02448-8.
- UN Inter-agency Group for Child Mortality Estimation. Levels and trends in child mortality: Report 2024, 2024. URL <https://data.unicef.org/resources/levels-and-trends-in-child-mortality-2024/>.
- World Health Organization, UNICEF, UNFPA, World Bank Group, and UNDESA/Population Division. Trends in maternal mortality 2000 to 2023, 2025. URL <https://www.who.int/publications/i/item/9789240108462>.
- Qing T. Zeng and Tony Tse. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1):24–29, 2006. doi: 10.1197/jamia.M1761.

A Vocabulary and rule examples

Table 2: Illustrative vocabulary entries. Patients describe the same symptom in English, in romanized Hindi, and code-mixed within a single phrase.

How patients say it	Symptom it maps to
bleeding from vagina	Vaginal bleeding
bleeding ho rahi hai	Vaginal bleeding
passing large clots	Vaginal bleeding
baby is not moving	Reduced fetal movement
bachcha move nahi kar raha	Reduced fetal movement
fits / jhatkay	Seizure
chest mein dard	Chest pain

Table 3: Example rules. Rules are scoped to a care category, so the same symptom is mapped to a different rule depending on what the patient is enrolled with us for: leg pain while she is pregnant (ANC) needs swelling and redness to be considered an emergency, while leg swelling post-delivery (PNC) becomes an emergency only if she has chest pain or breathing trouble. A *Standalone* rule applies on the symptom alone. A *Combination* rule applies only when a second symptom is present, or when the patient context matches, such as a given trimester or infant age.

Care category	Symptom	Also required	Emergency?
<i>Standalone</i>			
GHW	Seizure	nothing	Yes
HRPW	Fever above 38.5°C for 24 hours	nothing	Yes
<i>Combination: symptom + symptom</i>			
ANC	Burning urination	no fever	No
ANC	Burning urination	with fever	Yes
ANC	Leg pain	with leg swelling and redness	Yes
PNC	Leg swelling	with chest pain or breathing trouble	Yes
<i>Combination: symptom + context</i>			
ANC	Reduced fetal movement	trimester 1	No
ANC	Reduced fetal movement	trimester 2–3	Yes

B Dataset composition

Table 4: Distribution of the dataset by language

Language	N	%
Hindi	255	33.2
English	177	23.0
Telugu	152	19.8
Kannada	65	8.5
Punjabi	52	6.8
Marathi	46	6.0
Odia	22	2.9
Total	769	100

Table 5: Distribution of dataset by care category

Care category	N	%
ANC	377	49.0
PNC	175	22.8
HRPW	130	16.9
SNCU	28	3.6
GHW	20	2.6
MCH	15	2.0
GEMS	13	1.7
NCD	11	1.4
Total	769	100