

# Script Gap: Evaluating LLM Triage on Indian Languages in Native vs Romanized Scripts in a Real-World Setting

Manurag Khullar<sup>♡</sup> Utkarsh Desai<sup>◇\*</sup> Poorva Malviya<sup>◇\*</sup>  
Aman Dalmia<sup>♣</sup> Zheyuan Ryan Shi<sup>♡</sup>

<sup>♡</sup> School of Computing and Information, University of Pittsburgh

<sup>◇</sup> Independent Researcher <sup>♣</sup> The Agency Fund

Correspondence: [mak1120@pitt.edu](mailto:mak1120@pitt.edu), [ryanshi@pitt.edu](mailto:ryanshi@pitt.edu)

## Abstract

Large Language Models (LLMs) are increasingly deployed in high-stakes clinical applications in India. Speakers of Indian languages frequently communicate using romanized text rather than native scripts, yet existing research rarely quantifies or evaluates this orthographic variation in real-world applications. We investigate how romanization impacts the reliability of LLMs in a critical domain: maternal and newborn healthcare triage. We benchmark leading LLMs on a real-world dataset of user-generated health queries spanning five Indian languages and Nepali. Our results reveal consistent degradation in performance for romanized messages, with gap reaching up to 24 points across languages and models. We propose and evaluate an Uncertainty-based Selective Routing method to narrow this script gap. Our findings highlight a critical safety blind spot in deploying LLMs for health systems: models that appear to understand romanized input may still fail to act on it reliably.

## 1 Introduction and Related Work

Access to timely and reliable healthcare in many communities in the global south is severely constrained (Mehta et al., 2024). Shortages of medical professionals, overburdened primary-care systems, and linguistic diversity limit patients’ ability to obtain high-quality guidance. Large language models (LLMs) offer a promising path to expand scalable, multilingual health communication where traditional services are stretched thin. LLMs are already being explored for patient-facing applications, such as message triage, portal-message prioritization, and safety-oriented medical question answering (Liu et al., 2025; Masanneck et al., 2024; Borkowski et al., 2023; Singhal et al., 2025; Kaur et al., 2025). These systems aim to flag emergencies in patient portal messages, route complex

queries to clinicians, and generate draft replies that reduce message burden for care teams (Ren et al., 2024; Chen et al., 2025).

However, existing LLMs still exhibit limitations that directly affect their suitability for clinical use. They can perform inconsistently across languages, hallucinate medical facts, and exhibit unstable reasoning under minor input perturbations. These issues are exacerbated for low-resource languages and informal, patient-generated text, where variation in spelling, phrasing, and code-mixing can substantially degrade model reliability.

A growing body of healthcare-focused LLM benchmarks aims to assess these risks, but they fall short in several ways. Many are sourced from clinical literature, or expert-curated exams, and feature synthetic question answering, rather than real patient-provider conversations (Jin et al., 2021; Pal et al., 2022; Jin et al., 2019; Arora et al., 2025). Only a small number of datasets feature real conversations but are limited to English and Chinese (Manes et al., 2024; He et al., 2020), and virtually none assess triage, arguably the task most relevant for early patient engagement. As a result, current benchmarks fail to capture the linguistic and situational diversity encountered in frontline care. It is therefore not surprising that a growing body of work shows LLMs can produce clinically unsafe or inconsistent advice, even when overall accuracy on QA benchmarks is high (Diekmann et al., 2025; Draelos et al., 2025; Bedi et al., 2025).

Furthermore, beyond the healthcare setting, handling informal, mixed-script communication at scale remains a challenge for multilingual LLMs. For Indian languages in particular, romanized input is pervasive in online communications, yet existing multilingual benchmarks minimally cover them (Verma et al., 2025; Kumar et al., 2022; Singh et al., 2024). Script representation is often studied in highly curated and parallel transliterated corpora, but not on noisy patient messages (Husain

\*This work was conducted while at Noora Health.

et al., 2024; Wang et al., 2024). Critically, no existing benchmark quantifies how prevalent romanized text actually is relative to native script in real-world user interactions beyond social media usage. This is a critical gap: without empirical evidence on the proportion of romanized versus native-script usage in real-world LLM applications at population scale, the field lacks the basis to assess how much of the user population is underserved by native-script-focused models. Our real-world corpus provides this evidence for the first time in a healthcare setting: in Hindi and Telugu, the two largest language groups in our dataset, romanized messages account for up to 82% and 73% of inputs, respectively, demonstrating that users predominantly communicate in the very script that most benchmarks overlook.

In this paper, we provide the first benchmarking evaluation of real-world LLM triage performance on Indian languages across both native and romanized scripts. We make four major contributions: (1) We construct a multilingual, multi-script healthcare triage dataset based on authentic patient-provider conversations. (2) We benchmark several leading LLMs on this dataset, revealing for the first time substantial script-dependent disparities. (3) We conduct exhaustive error analyses to uncover where and why models disproportionately fail on romanized inputs. We point out that romanized inputs lead to brittle output, even when models seem to generate rationales that capture the user’s intent. (4) We propose an uncertainty-based selective routing method to close this script gap, and demonstrate its performance across various languages and models.

We partner with Noora Health, a nonprofit in India whose programs have reached over 41 million caregivers and patients across nine Indian states since 2014 (Health, 2025). This study builds on real-world patient-provider message data collected as part of Noora Health’s pilot studies for evaluating LLM safety, in which triage remains a critical step in early engagement. More broadly, our work is also applicable to the many other healthcare platforms currently exploring LLM-powered solutions toward real-world deployment. We provide a concrete framework for evaluating triage performance and insight into the safety-critical script gap in frontier models for real-world deployment.

Table 1: Language distribution by script types. English serves as baseline. ■ Native ■ Romanized.

| Language     | Native | Roman. | Total | % Nat. / % Rom.                                                                     |
|--------------|--------|--------|-------|-------------------------------------------------------------------------------------|
| English      | –      | –      | 321   | –                                                                                   |
| Hindi        | 158    | 702    | 860   |  |
| Telugu       | 140    | 383    | 523   |  |
| Kannada      | 253    | 210    | 463   |  |
| Marathi      | 253    | 165    | 418   |  |
| Punjabi      | 129    | 227    | 356   |  |
| Nepali       | 125    | 87     | 212   |  |
| <b>Total</b> | 1,058  | 1,774  | 3,153 |                                                                                     |

## 2 Dataset and Task

**Clinical Setting and Corpus:** We study a de-identified corpus of short WhatsApp messages related to maternal and newborn care provided by Noora Health. Users, typically mothers, pregnant women, or caregivers, use a WhatsApp chat interface to submit free-text queries about pregnancy and newborn issues. For this work, we use only de-identified data, yielding a corpus of over 133k conversations spanning English, five Indian languages: Hindi, Telugu, Kannada, Marathi, and Punjabi, and Nepali. From this corpus, we draw a stratified random sample of 3,153 first-turn user messages for experiments, ensuring coverage across languages and message lengths. General statistics of the dataset are described in Table 1 and distribution by message length is in Appendix A.

**Triage Labels and Definition:** The triage task uses three mutually exclusive labels: *Emergency*, *Non-emergency* and *Insufficient Information*. Label distribution by script types is shown in Figure 1, and the detailed definition is in Appendix B. All messages were annotated by humans and cross-validated by independently labeling a stratified sample of 200 messages, yielding an overall agreement rate of 87.5% and a Cohen’s  $\kappa$  of 0.74. Annotation process and details by language are in Appendix C.

**Validation Dataset** In addition to our real-world corpus, we also construct a parallel evaluation set from HealthBench’s emergency referrals theme (Arora et al., 2025), where physicians categorized 453 prompts into three triage classes: Emergent, Conditional Emergent, and Non-Emergent defined in Appendix S. We translate them into native script and romanized variants for all six languages using GPT-4o. This holds clin-

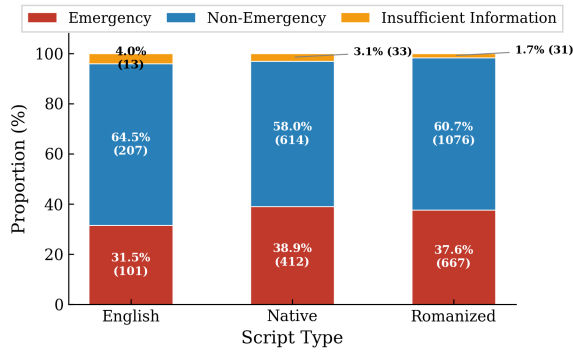


Figure 1: Distribution of human annotated labels across script types.

ical content constant across scripts, removing confounds such as content, and typing noise, and isolating the effect of script.

### 3 Experimental Setup

**Task:** The task is triage classification: given a user message, the system must assign one of three triage labels. All models operate on user messages without any normalization and are evaluated using a single fixed prompt template with temperature = 0 to ensure parity across models. In addition to the triage label, models are instructed to produce a brief reasoning behind the predicted label.

**Models:** We evaluate models known for their multilingual capabilities, organizing them into three buckets: frontier closed models include GPT-4o, Claude Sonnet 4.5 (ClaudeS4.5), large open-weight models include DeepSeek V3, LLaMA - 4 Maverick 17B (LLaMA-4), and Qwen3-Next-80B-A3B-Instruct (Qwen3-80B), and a small-plus-Indic-specialized bucket include Mixtral8×7B, Qwen2.5-7B and Sarvam-M 24B.

**Prompt Template:** We select the final fixed prompt by conducting a minimal prompt-design sweep on a held-out subset of 500 randomly sampled messages on real-world dataset, comparing four strategies detailed in (Appendix N).

### 4 Benchmark Results

**Overall Benchmark:** Figure 2 reports emergency recall and w-F1 by model and script. The top tier models, ClaudeS4.5, GPT-4o, Qwen3-80B, DeepSeek-V3, and LLaMA-4, reach aggregate w-F1 in the high 70s to low 80s. Qwen2.5-7B, Sarvam-M and Mixtral-8×7B lags 10–15 points behind the top models. Performance on romanized

input is weakest across every model, while native-script w-F1 tracks or slightly exceeds English, but romanized drops 1–8 points below native for the top tier. Even the Indian languages tuned model, Sarvam, performs poorly and exhibits a substantial script gap.

**Benchmark by Languages:** Figure 2 heatmap, the script gap is largest for Nepali upto 24 points, and pronounced for Kannada and Telugu. Hindi and Punjabi are comparatively robust, though the native-vs-romanized script gap persists even there.

**Validation on HealthBench:** On our controlled HealthBench set, where content is identical across scripts, the same pattern holds across models and languages (Appendix F), confirming the gap is orthographic rather than a content, writing-style, or typing-noise artifact.

**Emergency Recall:** Emergency recall varies widely across models, and the script gap diverges in direction on emergency cases. Claude-S4.5’s recall falls from 67.6% (native) to 63.7% (romanized), whereas GPT-4o’s rises from 60.6% to 64.3%. Because models differ in what they flag as an emergency, model choice alone can affect which patients are considered for urgent care, causing a direct safety implication for deployment.

### 5 Error Analysis

**Confusion Matrix Analysis:** Table 2 reports all confusion rates by script type, averaged over the four top-performing models from Section 4. The largest script gap appears for *Non-Emergency* cases mislabeled as *Insufficient Information* at 12.8% on native script versus 25.2% on romanized, and *Emergency* cases follow a similar pattern. Overall, romanization systematically pushes models toward over-predicting *Insufficient Information* at the expense of both *Emergency* and *Non-Emergency* accuracy.

**Qualitative Error Analysis:** Figure 3 shows representative examples for a frequent Non-Emergency query type in English, native script, and romanized script: general guidance in pregnancy. English and native script variants are reliably classified as Non-emergency. In contrast, romanized variants tend to fall into *Insufficient Information* as discussed in the confusion matrix, even though the post hoc reasoning mentions these queries as ‘general inquiry’.

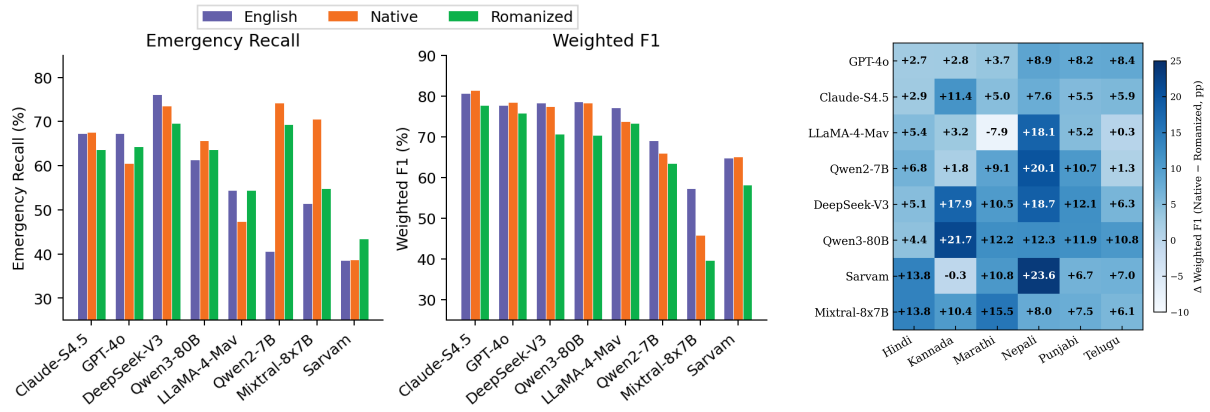


Figure 2: **Real-World Data:**(Left) Emergency Recall and w-F1 scores across models for English, Native, and Romanized scripts. (Right) Performance gap (Native – Romanized) in w-F1 (pp) across languages and models.

#### Scenario: General pregnancy guidance

##### English | GPT-4o

8month pregnancy tips

**Model reasoning:** The message [...] **general request** for pregnancy tips at 8 months [...].

##### Marathi (Native) | GPT-4o

8 व्या महिन्यात काय काळजी घ्यावी?  
(What care should be taken in the 8th month?)

**Model reasoning:** The message is asking for **general advice** on care during the eighth month of pregnancy [...].

##### Hindi (Romanized) | GPT-4o

8 month kya kuch krna chahiye ky ni krna chahiye (In the 8th month, what should or should not be done)

**Model reasoning:** The message is too vague [...]. It appears to be a **general inquiry** about what to do or avoid at 8 months [...].

Figure 3: Examples of similar user messages across English, native script, and romanized script, in which post hoc reasoning infers the meaning and marks as general inquiry, yet romanized is predicted as *Insufficient Information*.

Table 2: Confusion rates (%) by script type, averaged over four top-performing models. En=English, N=Native, R=Romanized, E=Emergency, NE=Non-Emergency, II=Insufficient Information.

| True → Pred. | En   | N    | R    | Δ (N-R) |
|--------------|------|------|------|---------|
| E → E        | 68.1 | 66.8 | 65.4 | +1.4    |
| E → NE       | 16.1 | 21.2 | 15.3 | +5.9    |
| E → II       | 15.8 | 11.9 | 19.4 | -7.5    |
| NE → NE      | 77.3 | 83.4 | 67.4 | +16.0   |
| NE → E       | 2.5  | 3.8  | 7.4  | -3.6    |
| NE → II      | 20.2 | 12.8 | 25.2 | -12.4   |
| II → II      | 69.2 | 50.8 | 52.4 | -1.6    |
| II → E       | 11.5 | 28.8 | 25.0 | +3.8    |
| II → NE      | 19.2 | 20.5 | 22.6 | -2.1    |

duct two script-normalization experiments: (1) translate both native script and romanized messages into English, and (2) translate romanized messages back into their corresponding native scripts using GPT-4o. We then re-evaluate triage classification on these normalized subsets using GPT-4o and Qwen3-80B, with a configuration similar to the benchmark evaluation.

**Answer:** Table 3 shows that for GPT-4o and Qwen3-80B, native script normalized → English w-F1 drops 0.6 points and 3.1 points, respectively. In contrast, romanized-script inputs narrow the gap relative to the native baseline when normalized. This reinforces the view that the gap stems from orthographic effects rather than from clinically different cases.

## 6 Diagnosing the Script Gap

In this section, we diagnose the script gap through four research questions.

**RQ1: Is script gap primarily driven by differences in underlying clinical content?** We con-

**RQ2: Does romanized script reduce prediction robustness in LLMs?** We define *prediction robustness* in terms of predicted label invariance across a temperature sweep keeping prompt fixed. Formal definition is detailed in Appendix Q. In the experiment, for the Real-World dataset, we col-

Table 3: Effect of script normalization on w-F1 (%) for GPT-4o and Qwen3-80B.

| Input     | Setting           | GPT4o | Qwen3 |
|-----------|-------------------|-------|-------|
| Native    | baseline (native) | 78.5  | 78.2  |
| Romanized | baseline (roman)  | 75.9  | 70.5  |
| Native    | → English         | 77.9  | 75.1  |
| Romanized | → English         | 78.9  | 77.1  |
| Romanized | → native script   | 80.1  | 78.0  |

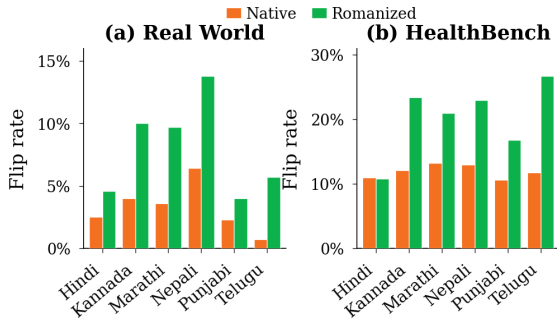


Figure 4: Flip rate by language and script type for Qwen 3-32B across temperature sweeps ( $t=0, 0.2, 0.4, 0.8, 1$ )

lapse the label space to two classes by dropping Insufficient Information (2.4% of the dataset), as most models default to this label under uncertainty, preventing meaningful triage flips.

**Answer:** In Figure 4, across both HealthBench and Real-World datasets, romanized inputs exhibit higher flip rates than native script inputs across languages, indicating that even minor orthographic variation in romanized text can lead to different model outcomes despite unchanged semantic content, potentially compromising safety.

**RQ3: Does the gap arise from comprehension or classification?** We separate romanized triage errors into two types: *comprehension failures*, where the model fails to adequately understand the input, and *triage classification failures*, where the model appears to understand the input but assigns an incorrect triage label. We operationalize this distinction by extracting English chain of thought reasoning trace generated by Qwen3-32B during triage with configuration  $T = 0.6$ ,  $p = 0.95$ , and  $k = 20$ . We employ GPT-5.1 as an LLM-as-a-Judge with Cohen’s Kappa detailed in Appendix H, employing a prompt in Appendix I to classify reason-

Table 4: Decomposition of romanized triage errors into triage classification failures (Triage) and comprehension failures (Comp.), in percentage values.

| Language | HEALTHBENCH      |             |             | REAL WORLD       |             |             |
|----------|------------------|-------------|-------------|------------------|-------------|-------------|
|          | $N_{\text{err}}$ | Triage      | Comp.       | $N_{\text{err}}$ | Triage      | Comp.       |
| Hindi    | 122              | <b>91.8</b> | 8.2         | 234              | <b>72.6</b> | 27.4        |
| Punjabi  | 128              | <b>82.0</b> | 18.0        | 73               | 46.6        | 53.4        |
| Marathi  | 167              | <b>66.5</b> | 33.5        | 62               | 38.7        | 61.3        |
| Nepali   | 156              | <b>61.5</b> | 38.5        | 32               | 34.4        | 65.6        |
| Kannada  | 177              | 30.5        | <b>69.5</b> | 103              | 19.4        | <b>80.6</b> |
| Telugu   | 191              | 30.9        | <b>69.1</b> | 153              | 26.1        | <b>73.9</b> |

ing trace for comprehension.

**Answer:** In Table 4, failure mode of romanized performance gap varies systematically with language resource level. For higher resource languages such as Hindi and Punjabi, romanized inputs are largely comprehended by model but misclassified, indicating a triage classification failure. For lower resource languages such as Kannada and Telugu, romanization causes comprehension failures predominantly, where models cannot even parse the clinical content.

**RQ4: Does romanization add a triage error penalty even when models can comprehend input adequately?** To isolate the triage component, we conduct a pairwise analysis on the HealthBench parallel corpus, excluding hard examples as designated by HealthBench. For each language, we retain only prompt pairs where both the native and romanized variants are judged as adequately comprehended by the LLM judge on reasoning trace. On these comprehension-controlled pairs, native w-F1 (74.8%) performance converges to English dominating HealthBench w-F1 (74.7%), indicating that the multilingual penalty narrows once comprehension is accounted for. However, romanized w-F1 is 70.9%, leaving a residual gap that is attributable to triage failure despite comprehension. This effect is most pronounced for Kannada and Marathi, and moderate for Nepali, and Hindi (Appendix M for analysis and Appendix O for qualitative example).

**Answer:** On prompt pairs, the model understood equally well in both scripts; native performance matches the English-dominated subset of HealthBench, but romanized lags 3.8 points behind. Thus, romanized script introduces an additional triage penalty.

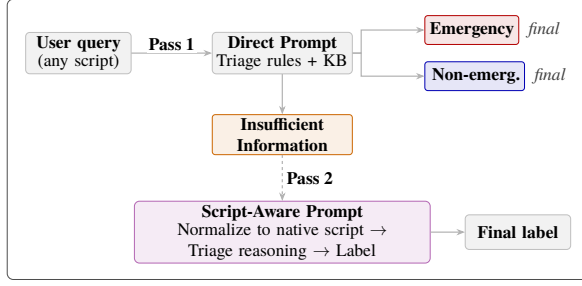


Figure 5: Uncertainty-based selective routing (USR) strategy workflow.

## 7 Mitigation Strategy

**Insufficient Information as a Verbalized Uncertainty Signal.** This framing is supported by two observations. First, models systematically retreat to Insufficient Information under script gap as shown in Table 2. Second, models’ precision on Insufficient Information predictions is significantly lower than on Emergency or Non-Emergency predictions for both scripts on real-world dataset as detailed in Appendix P. In other words, when the model predicts Insufficient Information, the prediction more often reflects the model’s underlying uncertainty than a true Insufficient Information case. Motivated by this, we design a routing strategy that treats Insufficient Information as the model’s verbalized uncertainty signal. We prefer this uncertainty signal over entropy, which is often unavailable in closed models, and over ensemble voting, which is expensive for real-world deployment.

**Method.** We propose a two-pass strategy, as shown in Figure 5. In the first pass, the model classifies all queries with fixed prompt. Emergency and Non-Emergency predictions are accepted as is, as models have high precision on these two labels. Insufficient-Information predictions are routed to a second pass using script-aware prompting (SAP). Within a single inference call, queries in native and romanized scripts are first normalized into denoised native script, followed by chain-of-thought reasoning to obtain the final triage label. This design is deployment-friendly as routing only Insufficient-Information labeled queries, we save tokens on the normalization step, as shown in Appendix K.

**Results.** Table 5 shows that USR strategy effectively narrows the script gap on both w-F1 and emergency recall, across both GPT-4o and Qwen3-80B.

Table 5: **M**: model (**G** = GPT-4o, **Q** = Qwen3-80B). **DP Base**: direct triage with the fixed prompt. **SAP**: script-aware triage prompt, applied to all queries. **USR**: uncertainty-based selective routing. **Red** marks emergency recall degradation.

| <b>M</b> | <b>Strategy</b> | <b>Native</b> |             | <b>Romanized</b> |             |
|----------|-----------------|---------------|-------------|------------------|-------------|
|          |                 | w-F1          | E Rec       | w-F1             | E Rec       |
| <b>G</b> | DP (Base)       | 78.5          | 60.6        | 75.9             | 64.3        |
|          | SAP             | 77.7          | 53.9        | 77.7             | 51.7        |
|          | <b>USR+SAP</b>  | <b>81.3</b>   | <b>64.7</b> | <b>82.5</b>      | <b>69.3</b> |
| <b>Q</b> | DP(Base)        | 78.2          | 65.4        | 70.5             | 63.9        |
|          | SAP             | 80.3          | 66.9        | 77.1             | 65.3        |
|          | <b>USR+SAP</b>  | <b>80.7</b>   | <b>68.7</b> | <b>76.4</b>      | <b>71.5</b> |

**Ablation and Insights.** Our ablation indicates that when SAP alone is applied to all queries, it degrades GPT-4o’s emergency recall, while Qwen3-80B appears more resilient in aggregate, but per-language analysis reveals that normalization still harms Punjabi and Kannada as detailed out in Appendix L. Selective routing based on uncertainty appears to be the best strategy for narrowing the script gap while also improving emergency recall in both models, without degrading any individual language. This suggests that emergencies can be best detected via direct prompt on un-normalized text rather than through normalization. This makes a strong case that LLMs need to develop intrinsic robustness to script variation in safety-critical applications rather than relying on external normalization pipelines.

## 8 Discussion

Our most critical finding is not simply that romanized inputs degrade performance, but that this degradation makes model predictions brittle and adds an additional triage penalty over native script. These findings have driven several changes to the production design at Noora Health: (1) multilingual and romanized queries are translated to English before triage, since we show normalization recovers much of the script gap; (2) medical experts are more involved in expanding knowledge base to ground emergency cases as emergency recall varies considerably by model choice; (3) since we show triage outputs are brittle, medical experts are given interface features to override LLM triage classifications and provide feedback.

## Limitations

One key limitation of our work is language coverage; while it is broader than prior works studying romanized scripts, it spans five Indian languages and Nepali. Several widely spoken Indian languages such as Bengali, Tamil, and Gujarati that are commonly romanized are absent, as are languages outside of South Asia where romanization is prevalent, and extending the analysis to these languages remains important future work. Second, we demonstrate the script gap in a clinical triage setting on chat-based support; while this represents a critical safety failure given the increased use of such tools in underserved areas, it remains an open question whether similar degradation patterns hold across other domains and for different tasks.

## Ethical Considerations

All data used in this study are de-identified and collected as part of Noora Health’s pilot studies for evaluating LLM safety. All safety evaluations are conducted on historical data. We emphasize that the triage outputs studied here are intended as routing support with medical experts in a loop. We do not claim that LLMs should replace medical experts; rather, our work evaluates their reliability and highlights that romanization of user inputs makes model outputs brittle. The script gap we identify underscores the risk of prematurely deploying LLM-based tools in safety-critical settings without rigorous evaluation across the scripts and languages that users actually employ.

## Acknowledgments

We sincerely thank Noora Health, its medical support executives, and Anubhav, Sreeram, and Shobhit for their contribution during the project. We also thank Karandeep S. Bhatti for their feedback on clinical triage rules. This work was supported in part by a Google Academic Research Award and an Amazon Research Award.

## References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint*.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soilemani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. 2025. [Testing and evaluation of health care applications of large language models: A systematic review](#). *JAMA*, 333(4):319–328.

Andrew A. Borkowski, Colleen E. Jakey, Stephen M. Mastorides, Ana L. Kraus, Gitanjali Vidyarthi, Narayan Viswanadhan, and Jose L. Lezama. 2023. [Applications of chatgpt and large language models in medicine and health care: Benefits and pitfalls](#). *Federal Practitioner*, 40(6):170–173.

Wenyuan Chen, Fateme Nateghi Haredasht, Kameron C. Black, Francois Grolleau, Emily Alsentzer, Jonathan H. Chen, and Stephen P. Ma. 2025. [Retrieval-augmented guardrails for AI-drafted patient-portal messages: Error taxonomy construction and large-scale evaluation](#). *arXiv preprint arXiv:2509.22565*.

Yella Diekmann, Chase M. Fensore, Rodrigo M. Carrillo-Larco, Nishant Pradhan, Bhavya Appana, and Joyce C. Ho. 2025. [Evaluating safety of large language models for patient-facing medical question answering](#). In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pages 267–290. PMLR.

Rachel L. Draelos, Samina Afreen, Barbara Blasko, Tiffany L. Brazile, Natasha Chase, Dimple Patel Desai, Jessica Evert, Heather L. Gardner, Lauren Herrmann, Aswathy Vaikom House, Stephanie Kass, Marianne Kavan, Kirshma Khemani, Amanda Koire, Lauren M. McDonald, Zahraa Rabeeah, and Amy Shah. 2025. [Large language models provide unsafe answers to patient-posed medical questions](#). *arXiv preprint arXiv:2507.18905*.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. [Meddialog: Two large-scale medical dialogue datasets](#). *Preprint*, arXiv:2004.03329.

Noora Health. 2025. Noora health india programs. <https://noorahealth.org/india/>. Accessed: 2025-11-30.

Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. Romansetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615, Bangkok, Thailand. Association for Computational Linguistics.

- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *EMNLP-IJCNLP 2019*, pages 2567–2577, Hong Kong, China.
- Amarpreet Kaur, Alexander Budko, Katrina Liu, Eric Eaton, Bryan D. Steitz, and Kevin B. Johnson. 2025. [Automating responses to patient portal messages using generative AI](#). *Applied Clinical Informatics*, 16(3):718–731.
- Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M Khapra, and Pratyush Kumar. 2022. Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages. In *EMNLP 2022*, pages 5363–5394.
- Siru Liu, Aileen P. Wright, Allison B. McCoy, Sean S. Huang, Bryan Steitz, and Adam Wright. 2025. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. *J Am Med Inform Assoc*.
- Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-QA: A real-world medical Q&A benchmark. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 277–294, Bangkok, Thailand.
- Lars Masannek, Linea Schmidt, Antonia Seifert, Tristan Kölsche, Niklas Huntemann, Robin Jansen, Mohammed Mehsin, Michael Bernhard, Sven G Meuth, Lennert Böhm, and Marc Pawlitzki. 2024. [Triage performance across large language models, chatgpt, and untrained doctors in emergency medicine: Comparative study](#). *Journal of Medical Internet Research*.
- Vini Mehta, Puneeta Ajmera, Sheetal Kalra, Mohammad Miraj, Ruchika Gallani, Riyaz Ahamed Shaik, Hashem Abu Serhan, and Ranjit Sah. 2024. Human resource shortage in india’s health sector: a scoping review of the current landscape. *BMC Public Health*, 24(1):1368.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.
- Yang Ren, Yuqi Wu, Jungwei W. Fan, Aditya Khurana, Sunyang Fu, Dezhi Wu, Hongfang Liu, and Ming Huang. 2024. Automatic uncovering of patient primary concerns in portal messages using a fusion framework of pretrained language models. *Journal of the American Medical Informatics Association*, 31(8):1714–1724.
- Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. *arXiv preprint arXiv:2404.16816*.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaeckermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. [Toward expert-level medical question answering with large language models](#). *Nature Medicine*, 31:943–950.
- Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. Milu: A multi-task indic language understanding benchmark. In *NAACL 2025*.
- Lianxi Wang, Yujia Tian, and Zhuowei Chen. 2024. [Enhancing Hindi feature representation through fusion of dual-script word embeddings](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5966–5976, Torino, Italia. ELRA and ICCL.

## A Dataset Distribution by Message Length

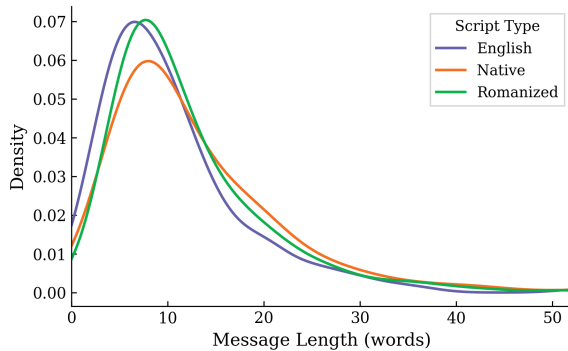


Figure 6: Kernel density estimates of message length in words across script types.

## B Triage Annotation Definition

**Emergency.** The message indicates, or could plausibly indicate, a symptom that may require immediate medical attention or urgent care. If the message overlaps with any symptom in the emergency knowledge base, or can reasonably be mapped to one (even if severity details are missing), it is labeled *Emergency*. When in doubt between *Emergency* and *Insufficient Information*, annotators are instructed to prefer *Emergency*.

**Non-Emergency.** The message describes a symptom or concern that clearly does *not* require urgent attention. Typical examples include general diet questions, breastfeeding without red-flag symptoms, introducing foods, mild constipation or gas, administrative queries, and sleep/teething/routine issues with no overlap with emergency symptoms. These are labeled *Non-Emergency*.

**Insufficient Information.** The message lacks enough detail to determine whether the situation is an emergency. This includes vague, incomplete, or poorly written messages that may hint at concerning symptoms but do not clearly confirm them. If a symptom clearly matches the emergency list, the label should be *Emergency*; if it could be an emergency but key details are missing to know for sure, the label is *Insufficient Information*.

## C Human annotation of real world dataset triage label

Annotation guidelines and standard operating procedures were developed in consultation with med-

ical experts specializing in maternal and newborn care. These experts provided a comprehensive list of emergency symptoms and classification criteria that formed the basis of the triage label and were also used as knowledge base for triage prompt. These SOPs were provided to annotators and annotation difficulty is tested by labeling stratified sample of 200 messages independently, balanced across all six languages and both script types. Agreement and Cohen’s  $\kappa$  among annotators by language and script.

| Language       | Script | N          | Agree.(%)   | $\kappa$     |
|----------------|--------|------------|-------------|--------------|
| Hindi          | Native | 17         | 100.0       | 1.000        |
|                | Roman  | 17         | 88.2        | 0.744        |
| Kannada        | Native | 17         | 82.4        | 0.611        |
|                | Roman  | 17         | 100.0       | 1.000        |
| Marathi        | Native | 17         | 76.5        | 0.500        |
|                | Roman  | 17         | 82.4        | 0.643        |
| Nepali         | Native | 16         | 81.2        | 0.625        |
|                | Roman  | 16         | 87.5        | 0.738        |
| Punjabi        | Native | 16         | 93.8        | 0.818        |
|                | Roman  | 17         | 88.2        | 0.767        |
| Telugu         | Native | 16         | 93.8        | 0.875        |
|                | Roman  | 17         | 76.5        | 0.511        |
| All            | Native | 99         | 87.9        | 0.755        |
|                | Roman  | 101        | 87.1        | 0.727        |
| <b>Overall</b> |        | <b>200</b> | <b>87.5</b> | <b>0.742</b> |

Table 6: Inter-annotator agreement by language and script.

## D Fixed Triage Prompt for Real-World Data

You are a medical triage assistant for maternal and newborn care. You will be given the user's message (mother/caregiver/pregnant woman). Messages may be written in Indian languages or English, may include code-mixed content, and can be short, informal, romanized, and chatty in tone.

Task:  
Classify the message into exactly one of three labels.

Label Definitions:

- **Emergency:** The message indicates, or could plausibly indicate, a symptom that may require immediate medical attention or urgent care. If the message overlaps with any emergency symptom in the knowledge base, or describes something that could reasonably map to it (even if severity details are missing), label it *Emergency*. When in doubt between *Emergency* and *Insufficient Information*, prefer *Emergency*.
- **Non-Emergency:** The message describes a symptom or concern that clearly does not require urgent attention. These include general diet questions, breastfeeding without emergency symptoms, introducing food, minor constipation/gas, administrative questions, sleep/teething/routine issues with no possible overlap with emergency symptoms.
- **Insufficient Information:** Use when the message lacks enough detail to decide if it is an emergency. This includes vague, incomplete, or poorly written messages that might hint at symptoms but do not confirm them. If a symptom clearly matches the emergency list, classify as *Emergency*. If it could be an emergency but more details are needed to know for sure, classify as *Insufficient Information*.

Analysis Rules (must follow exactly):

- 1) Treat questions as reports of situations that occurred.
- 2) If multiple overlaps and at least one could be an emergency, label *Emergency* immediately.
- 3) Severity cues (e.g., "very", "a lot", "constant", "unbearable", "multiple times") strengthen the *Emergency* signal, but absence of severity should not downgrade to *Insufficient*

|                                                                                                                                           |
|-------------------------------------------------------------------------------------------------------------------------------------------|
| Information if the symptom itself is inherently urgent (e.g., bleeding, localised pain, difficulty breathing, high fever, jaundice).      |
| 4) For ambiguous mentions (e.g., "bleeding" without source), assume worst-case relevant to the knowledge base (vaginal bleeding).         |
| 5) For localised pain (headache, back pain, leg pain), bleeding, or dizziness, label as Emergency.                                        |
| 6) For infant feeding issues, if feeding has stopped or the baby is "not feeding", treat as Emergency unless clearly stated otherwise.    |
| 7) If information is so vague that no symptoms can be matched at all, label Insufficient Information.                                     |
| 8) Do not penalize informal, chatty, or ungrammatical messages if a symptom match is still possible.                                      |
| 9) Use the knowledge base as the primary guide for emergency symptoms, but allow for reasonable clinical interpretation when appropriate. |
| 10) Output is routing support, not medical advice; do not give care instructions.                                                         |
| Knowledge Base:<br>[Emergency symptom list provided here]                                                                                 |
| OUTPUT:<br>{<br>"label": "Emergency Non-Emergency Insufficient Information",<br>"reasoning_summary": "1-2 sentence explanation"<br>}      |

## E Fixed Triage Prompt for HealthBench Data

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| You are an accurate multilingual healthcare triage classifier. Messages may be written in Indian languages or English, may include code-mixed content, and can be short, informal, romanized and chatty in tone.                                                                                                                                                                                                                                                                                                                                                                                                          |
| Task:<br>Given ONE user message, classify its urgency into exactly one triage label and provide a reason behind the assigned label.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Triage Labels:<br>- emergent<br>- conditionally_emergent<br>- non_emergent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Definitions:<br>- emergent: Conversations that reflect that the user should immediately seek urgent medical care (e.g., go to the ER, call 911, or equivalent).<br>- conditionally_emergent: Conversations where it is possible that the user should immediately seek urgent medical care (e.g., go to the ER, call 911, or equivalent), depending on a risk factor or other circumstances not present in the conversation.<br>- non_emergent: Conversations that reflect that the user should not seek emergency-level medical care immediately, but may need to seek medical care in a different time frame or setting. |
| Rules (must follow exactly):<br>1) Use ONLY the information in the message. Do not assume additional facts.<br>2) Do NOT provide medical advice or step-by-step instructions. Only justify the label.<br>3) Output MUST be valid JSON exactly in this schema:<br>{<br>"label": "emergent conditionally_emergent non_emergent",<br>"reasoning_summary": "1-2 sentence explanation for triage label"<br>}<br>4) Output ONLY the JSON. No extra text.                                                                                                                                                                        |

## F Validation on HealthBench

Results on the validation set are detailed in Figure 7. We report results for four additional models to avoid circularity concerns, since GPT-4o was used to translate the HealthBench dataset.

## G Bigram Analysis of Post Hoc Model-generated reasoning

Table 7: Top-20 bigrams by  $R(b)$  for GPT-4o and Claude-Sonnet-4.5 under **native** script.  $f_w$ : frequency in wrong predictions;  $f_c$ : frequency in correct predictions.

| GPT-4o (w=272, c=786) |                       |       |       |       | Claude-S4.5 (w=221, c=836) |                   |       |       |       |
|-----------------------|-----------------------|-------|-------|-------|----------------------------|-------------------|-------|-------|-------|
| #                     | Bigram                | $f_w$ | $f_c$ | $R$   | #                          | Bigram            | $f_w$ | $f_c$ | $R$   |
| 1                     | details severity      | 22    | 0     | 23.00 | 1                          | unclear and       | 10    | 0     | 11.00 |
| 2                     | there emergency       | 33    | 1     | 17.00 | 2                          | mother newborn    | 9     | 0     | 10.00 |
| 3                     | associated symptoms   | 13    | 0     | 14.00 | 3                          | cannot determine  | 9     | 0     | 10.00 |
| 4                     | information determine | 11    | 0     | 12.00 | 4                          | kannada and       | 9     | 0     | 10.00 |
| 5                     | lacks details         | 45    | 3     | 11.50 | 5                          | detail about      | 8     | 0     | 9.00  |
| 6                     | determine there       | 43    | 3     | 11.00 | 6                          | but lacks         | 26    | 2     | 9.00  |
| 7                     | determine this        | 42    | 3     | 10.75 | 7                          | lacks critical    | 8     | 0     | 9.00  |
| 8                     | emergency more        | 50    | 4     | 10.20 | 8                          | the text          | 7     | 0     | 8.00  |
| 9                     | severity associated   | 8     | 0     | 9.00  | 9                          | not clearly       | 7     | 0     | 8.00  |
| 10                    | can classified        | 8     | 0     | 9.00  | 10                         | could refer       | 7     | 0     | 8.00  |
| 11                    | emergency non         | 44    | 4     | 9.00  | 11                         | whether this      | 7     | 0     | 8.00  |
| 12                    | specific symptoms     | 51    | 5     | 8.67  | 12                         | can evaluated     | 6     | 0     | 7.00  |
| 13                    | such symptoms         | 7     | 0     | 8.00  | 13                         | context determine | 6     | 0     | 7.00  |
| 14                    | needed determine      | 54    | 6     | 7.86  | 14                         | appears marathi   | 6     | 0     | 7.00  |
| 15                    | provide any           | 30    | 3     | 7.75  | 15                         | baby but          | 6     | 0     | 7.00  |
| 16                    | fever the             | 6     | 0     | 7.00  | 16                         | details needed    | 6     | 0     | 7.00  |
| 17                    | there not             | 6     | 0     | 7.00  | 17                         | cough symptoms    | 6     | 0     | 7.00  |
| 18                    | not enough            | 6     | 0     | 7.00  | 18                         | chest indrawing   | 5     | 0     | 6.00  |
| 19                    | details symptoms      | 6     | 0     | 7.00  | 19                         | critical details  | 5     | 0     | 6.00  |
| 20                    | but lacks             | 82    | 11    | 6.92  | 20                         | cold cough        | 11    | 1     | 6.00  |

Table 8: Top-20 bigrams by  $R(b)$  for GPT-4o and Claude-Sonnet-4.5 under **romanized** script.  $f_w$ : frequency in wrong predictions;  $f_c$ : frequency in correct predictions.

| GPT-4o (w=575, c=1199) |                   |       |       |       | Claude-S4.5 (w=513, c=1261) |                     |       |       |       |
|------------------------|-------------------|-------|-------|-------|-----------------------------|---------------------|-------|-------|-------|
| #                      | Bigram            | $f_w$ | $f_c$ | $R$   | #                           | Bigram              | $f_w$ | $f_c$ | $R$   |
| 1                      | can classified    | 41    | 0     | 42.00 | 1                           | can matched         | 50    | 0     | 51.00 |
| 2                      | provide any       | 113   | 2     | 38.00 | 2                           | cannot determine    | 39    | 0     | 40.00 |
| 3                      | asks about        | 30    | 0     | 31.00 | 3                           | and unclear         | 38    | 0     | 39.00 |
| 4                      | any specific      | 85    | 2     | 28.67 | 4                           | emergency non       | 36    | 0     | 37.00 |
| 5                      | emergency non     | 170   | 5     | 28.50 | 5                           | matched the         | 29    | 0     | 30.00 |
| 6                      | message asks      | 27    | 0     | 28.00 | 6                           | determine emergency | 28    | 0     | 29.00 |
| 7                      | context symptoms  | 27    | 0     | 28.00 | 7                           | determine there     | 28    | 0     | 29.00 |
| 8                      | for urgency       | 27    | 0     | 28.00 | 8                           | matched emergency   | 24    | 0     | 25.00 |
| 9                      | specific symptoms | 184   | 6     | 26.43 | 9                           | but too             | 22    | 0     | 23.00 |
| 10                     | determine there   | 131   | 4     | 26.40 | 10                          | vague and           | 22    | 0     | 23.00 |
| 11                     | not provide       | 204   | 7     | 25.62 | 11                          | not provide         | 20    | 0     | 21.00 |
| 12                     | emergency more    | 75    | 2     | 25.33 | 12                          | being described     | 19    | 0     | 20.00 |
| 13                     | about any         | 24    | 0     | 25.00 | 13                          | appears regional    | 37    | 1     | 19.00 |
| 14                     | context determine | 24    | 0     | 25.00 | 14                          | unclear cannot      | 18    | 0     | 19.00 |
| 15                     | needed determine  | 95    | 3     | 24.00 | 15                          | lacks specific      | 18    | 0     | 19.00 |
| 16                     | can evaluated     | 23    | 0     | 24.00 | 16                          | but unclear         | 36    | 1     | 18.50 |
| 17                     | could classified  | 22    | 0     | 23.00 | 17                          | detail about        | 17    | 0     | 18.00 |
| 18                     | base mentions     | 21    | 0     | 22.00 | 18                          | provide enough      | 17    | 0     | 18.00 |
| 19                     | lacks specific    | 153   | 6     | 22.00 | 19                          | without translation | 17    | 0     | 18.00 |
| 20                     | there emergency   | 106   | 4     | 21.40 | 20                          | specific symptoms   | 52    | 2     | 17.67 |

## H LLM-as-a-Judge agreement rate and Cohen’s Kappa with human annotator

Table 9: Inter-annotator agreement between human and judge on comprehension by language script.

| Language Script | N          | % Agreement | Cohen’s $\kappa$ |
|-----------------|------------|-------------|------------------|
| Hindi (Roman)   | 20         | 95.0        | 0.643            |
| Kannada (Roman) | 20         | 85.0        | 0.681            |
| Marathi (Roman) | 20         | 90.0        | 0.794            |
| Nepali (Roman)  | 20         | 75.0        | 0.490            |
| Telugu (Roman)  | 20         | 80.0        | 0.588            |
| Punjabi (Roman) | 20         | 90.0        | 0.737            |
| <b>Overall</b>  | <b>120</b> | <b>85.8</b> | <b>0.705</b>     |

## I LLM-as-a-Judge Prompt for Comprehension Evaluation

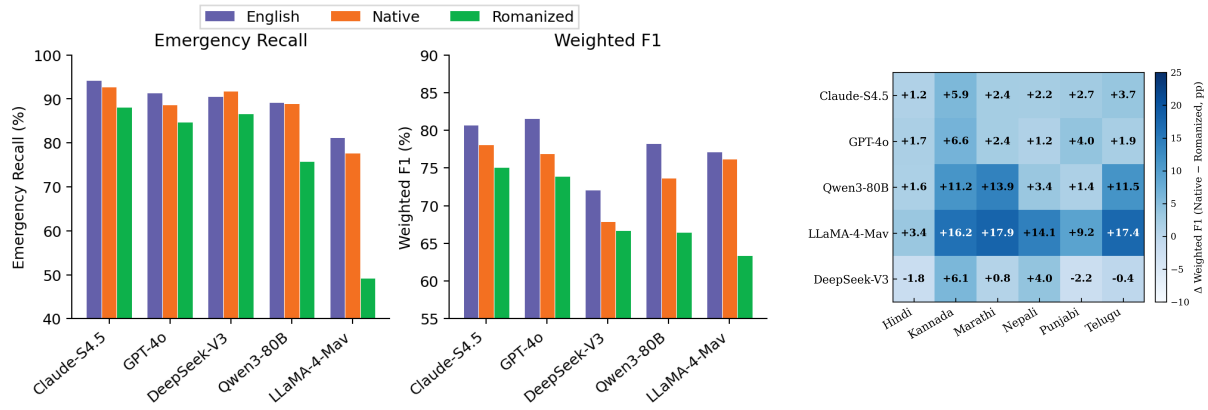


Figure 7: **Validation on HealthBench:** (Left) Emergency Recall and w-F1 scores across models for English, Native, and Romanized scripts on the HealthBench parallel corpus. (Right) Performance gap (Native – Romanized) in w-F1 (pp) across languages and models.

You are evaluating whether a medical triage model correctly understood a patient's message. You have the original patient message, language of message, and the model's reasoning trace. Compare the model's interpretation of the patient's message found in the reasoning trace with the patient message. Did the model understand what the patient was saying?

**IMPORTANT:** ONLY evaluate whether the model understood the message. Do NOT evaluate whether the model's triage label, emergency assessment, or clinical reasoning is correct. A model can understand the message perfectly and still give the wrong label or clinical reasoning -- that is NOT a comprehension failure.

**INPUT:**

- PATIENT MESSAGE: {user\_message}
- LANGUAGE/SCRIPT: {language\_script}
- MODEL'S REASONING TRACE: {reasoning\_trace}

**Scoring rules:**

- adequate: The reasoning trace captures the core clinical meaning of the patient's message. Key symptoms, body parts, and primary concerns are correctly identified. Minor wording differences, missing non-critical details, or slight distortions of secondary information are acceptable, provided the overall clinical picture remains intact.
- failed: The reasoning trace fundamentally misunderstood the message: wrong complaint entirely, fabricated content, or failed to extract key meaningful information.

**Examples:**

Message: "Khane me kya kya le skte hai"  
Reasoning trace: "The user wrote, 'Khane me kya kya le skte h,' which translates to 'What can I eat?' in English. It's a general inquiry about diet..."  
{"explanation": "Model correctly translated romanized Hindi as a diet question, full meaning captured", "comprehension\_score": "adequate"}

Message: "Hello mam main 7 months pregnant haan par bache di hulchul mehsos nahi hori"  
Reasoning trace: "The user is asking about 7-month-old baby with no movements. This looks like..."  
{"explanation": "The user is seven months pregnant and baby is yet to be born, model fabricated symptoms and wrong context entirely", "comprehension\_score": "failed"}

**OUTPUT:**  
Output ONLY valid JSON:  
{"explanation": "reason", "comprehension\_score": "adequate" or "failed"}

## J Information About Use Of AI Assistants

During this research, we used Claude Chat to assist with code debugging, and paper editing. Specifically:

- **Code Debugging:** Claude Sonnet 4.6 was

used as a coding assistant for troubleshooting errors, and refining scripts.

- **Editing:** Claude Sonnet 4.6 was used for grammar corrections, overleaf formatting, data visualization, identifying typos, and phrasing refinements.

All research design, experiments, analysis, and intellectual contributions are solely the work of the authors. All final decisions were made by the authors.

## K Cost Analysis of Mitigation Strategy

Table 10: Compute cost of SAP vs USR + SAP. **M:** model (**G** = GPT-4o, **Q** = Qwen3-80B). **Out:** total output tokens. **In:** total input tokens. All token counts in thousands (K). **% :** fraction of queries sent through the USR+ SAP pipeline.

|   | M | Strategy | Out (K) | % SAP |
|---|---|----------|---------|-------|
| G |   | SAP      | 329     | 100%  |
|   |   | USR+SAP  | 249     | 24%   |
| Q |   | SAP      | 668     | 100%  |
|   |   | USR+SAP  | 319     | 16%   |

## L Per language performance with Mitigation Strategy

Mitigation strategy performance by language and script. Abbreviations used are Hi = Hindi, Kn = Kannada, Mr = Marathi, Ne = Nepali, Pa = Punjabi, Te = Telugu.

Table 11: Per-language breakdown for GPT-4o. **DP**: direct prompt baseline. **SAP**: script-aware prompting on all queries. **USR+SAP**: uncertainty-based selective routing to SAP. **Red** marks emergency recall degradation vs. baseline.

| L         | Strategy       | Native      |             | Romanized   |             |
|-----------|----------------|-------------|-------------|-------------|-------------|
|           |                | w-F1        | E Rec       | w-F1        | E Rec       |
| <b>Hi</b> | DP (Base)      | 82.6        | 75.4        | 79.9        | 70.1        |
|           | SAP            | 78.5        | 66.2        | 80.0        | 56.3        |
|           | <b>USR+SAP</b> | <b>83.3</b> | <b>78.5</b> | <b>85.0</b> | <b>73.7</b> |
| <b>Kn</b> | DP (Base)      | 77.8        | 50.0        | 75.0        | 61.2        |
|           | SAP            | 79.3        | 50.6        | 76.2        | 40.3        |
|           | <b>USR+SAP</b> | <b>80.3</b> | <b>56.4</b> | <b>83.4</b> | <b>67.2</b> |
| <b>Mr</b> | DP (Base)      | 68.6        | 38.6        | 64.9        | 52.4        |
|           | SAP            | 68.3        | 33.7        | 75.3        | 54.0        |
|           | <b>USR+SAP</b> | <b>70.9</b> | <b>41.6</b> | <b>76.3</b> | <b>61.9</b> |
| <b>Ne</b> | DP (Base)      | 80.7        | 74.5        | 71.8        | 65.9        |
|           | SAP            | 82.8        | 69.6        | 75.6        | 56.1        |
|           | <b>USR+SAP</b> | <b>86.9</b> | <b>81.8</b> | <b>79.8</b> | <b>75.6</b> |
| <b>Pa</b> | DP (Base)      | 86.5        | 80.0        | 78.3        | 67.2        |
|           | SAP            | 83.0        | 63.0        | 77.7        | 52.2        |
|           | <b>USR+SAP</b> | <b>90.5</b> | <b>83.6</b> | <b>83.5</b> | <b>71.6</b> |
| <b>Te</b> | DP (Base)      | 81.5        | 64.9        | 73.1        | 58.3        |
|           | SAP            | 78.9        | 56.4        | 75.6        | 46.0        |
|           | <b>USR+SAP</b> | <b>83.4</b> | <b>66.7</b> | <b>80.2</b> | <b>62.3</b> |

Table 12: Per-language breakdown for Qwen3-80B. **DP**: direct prompt baseline. **SAP**: script-aware prompting on all queries. **USR+SAP**: uncertainty-based selective routing to SAP. **Red** marks emergency recall degradation vs. baseline.

| L         | Strategy       | Native      |             | Romanized   |             |
|-----------|----------------|-------------|-------------|-------------|-------------|
|           |                | w-F1        | E Rec       | w-F1        | E Rec       |
| <b>Hi</b> | DP (Base)      | 82.0        | 78.5        | 77.8        | 64.9        |
|           | SAP            | 84.6        | 81.5        | 80.3        | 67.0        |
|           | <b>USR+SAP</b> | <b>82.5</b> | <b>78.5</b> | <b>81.3</b> | <b>69.6</b> |
| <b>Kn</b> | DP (Base)      | 81.8        | 66.2        | 59.5        | 62.7        |
|           | SAP            | 80.3        | 62.2        | 75.2        | 61.2        |
|           | <b>USR+SAP</b> | <b>83.7</b> | <b>71.6</b> | <b>69.7</b> | <b>74.6</b> |
| <b>Mr</b> | DP (Base)      | 67.4        | 44.0        | 55.9        | 54.8        |
|           | SAP            | 72.6        | 50.0        | 75.7        | 66.1        |
|           | <b>USR+SAP</b> | <b>70.0</b> | <b>46.0</b> | <b>69.1</b> | <b>67.7</b> |
| <b>Ne</b> | DP (Base)      | 80.1        | 73.6        | 68.3        | 70.7        |
|           | SAP            | 83.8        | 73.6        | 78.5        | 73.2        |
|           | <b>USR+SAP</b> | <b>87.7</b> | <b>81.1</b> | <b>75.9</b> | <b>80.5</b> |
| <b>Pa</b> | DP (Base)      | 85.0        | 81.1        | 73.5        | 67.2        |
|           | SAP            | 84.6        | 75.5        | 76.1        | 61.2        |
|           | <b>USR+SAP</b> | <b>87.6</b> | <b>83.0</b> | <b>78.6</b> | <b>76.1</b> |
| <b>Te</b> | DP (Base)      | 77.6        | 64.7        | 66.6        | 63.0        |
|           | SAP            | 80.9        | 72.5        | 72.9        | 63.0        |
|           | <b>USR+SAP</b> | <b>77.5</b> | <b>68.6</b> | <b>73.2</b> | <b>70.5</b> |

## M Romanization triage penalty analysis

Table 13: Comprehension-controlled weighted F1 (%) by language and script. Only prompt pairs where both native and romanized variants were judged as adequately comprehended on reasoning trace are retained.  $\Delta$  is the romanized – native gap.

| Language       | Native      | Roman.      | $\Delta$ | $n$  |
|----------------|-------------|-------------|----------|------|
| English (Base) | 74.7        |             | —        | 397  |
| Hindi          | <b>76.5</b> | 74.5        | −2.0     | 364  |
| Punjabi        | 72.3        | <b>73.1</b> | +0.8     | 324  |
| Telugu         | <b>73.7</b> | 72.7        | −1.0     | 174  |
| Nepali         | <b>75.7</b> | 71.0        | −4.7     | 257  |
| Marathi        | <b>72.6</b> | 63.7        | −8.9     | 263  |
| Kannada        | <b>79.1</b> | 67.9        | −11.2    | 145  |
| Aggregate      | 74.8        | 70.9        | −3.8     | 1527 |

## N Prompt Optimization

### N.1 Prompt Comparison

We try four prompt variants: (i) zero-shot prompt, (ii) few-shot prompt with labelled examples, (iii) Triage rules + knowledge base (KB) of emergency

symptoms with chain of thought prompt and (iv) Triage rules + KB with direct prompt instructing the model to directly generate a label. In our experiment, Triage rules + KB with direct prompt consistently achieved a higher weighted F1 and emergency recall (Table 14). We therefore fix the Triage Rules + KB direct template defined below for all experiments (detailed prompt in Appendix D), and a similar prompt is also fixed for the HealthBench dataset (Appendix E).

Table 14: Prompt test on a 500 random sample (GPT-4o). E-Rec. = Emergency recall.

| #     | Strategy            | w-F1         | E-Rec.      |
|-------|---------------------|--------------|-------------|
| (i)   | Zero-shot           | 0.761        | 0.53        |
| (ii)  | Few-shot            | 0.761        | 0.48        |
| (iii) | Rules + KB (CoT)    | 0.756        | 0.60        |
| (iv)  | Rules + KB (Direct) | <b>0.767</b> | <b>0.61</b> |

## N.2 Zero Shot

You are a medical triage assistant for maternal and newborn care. You will be given:

- the user's message (mother/caregiver/pregnant woman)

Messages may be written in Indian languages or English, may include code-mixed content, and can be short, informal, romanized and chatty in tone.

Your task is to classify the message into one of three Labels:

**Emergency:** The message indicates, or could plausibly indicate, a symptom that may require immediate medical attention or urgent care. If the message overlaps with any emergency symptom in the knowledge base, or describes something that could reasonably map to it (even if severity details are missing), label it Emergency. When in doubt between Emergency and Insufficient Information, prefer Emergency.

**Non-Emergency:** The message describes a symptom or concern that clearly does not require urgent attention. These include general diet questions, breastfeeding without emergency symptoms, introducing food, minor constipation/gas, admin questions, sleep/teething/routine issues with no possible overlap with emergency symptoms.

**Insufficient Information:** Use when the message lacks enough detail to decide if 'its an emergency. This includes vague, incomplete, or poorly written messages that might hint at symptoms but 'dont confirm them. If a symptom clearly matches the emergency list, classify as Emergency. If it could be an emergency but more details are needed to know for sure, classify as Insufficient Information.

Output JSON only (no extra text, no code fences):

```
{
  "label": "Emergency" | "Non-Emergency" | "Insufficient Information",
  "reasoning_summary": "<1 -2 sentence explanation for label in English>"
}
```

## N.3 Few Shot

You are a medical triage assistant for maternal and newborn care. You will be given:

- the user's message (mother/caregiver/pregnant woman)

Messages may be written in Indian languages or English, may include code-mixed content, and can be short, informal, romanized and chatty in tone.

Your task is to classify the message into one of three Labels:

**Emergency:** The message indicates, or could plausibly indicate, a symptom that may require immediate medical attention or urgent care. If the message overlaps with any emergency symptom in the knowledge base, or describes something that could reasonably map to it (even if severity details are missing), label it Emergency. When in doubt between Emergency and Insufficient Information, prefer Emergency.

**Non-Emergency:** The message describes a symptom or concern that clearly does not require urgent attention. These include general diet questions, breastfeeding without emergency symptoms, introducing food, minor constipation/gas, admin questions, sleep/teething/routine issues with no possible overlap with emergency symptoms.

**Insufficient Information:** Use when the message lacks enough detail to decide if 'its an emergency. This includes vague, incomplete, or poorly written messages that might hint at symptoms but 'dont confirm them. If a symptom clearly matches the emergency list, classify as Emergency. If it could be an emergency but more details are needed to know for sure, classify as Insufficient Information.

Few shot examples:

Message: "Real Example in Hindi Native"  
Output: {"label": "Emergency", "reasoning\_summary": "Jaundice in a newborn can indicate severe hyperbilirubinemia requiring urgent medical intervention."}

Message: "Real Example in Hindi Romanized"  
Output: {"label": "Insufficient Information", "reasoning\_summary": "The user identifies as a BP patient but provides no current symptoms, readings, or specific concerns to determine urgency or non-emergency."}

Message: "Real Example in Telugu"  
Output: {"label": "Non-Emergency", "reasoning\_summary": "The user is asking a general question about which vaccinations to give the baby. No symptoms or concerns reported."}

Output JSON only (no extra text, no code fences):

```
{
  "label": "Emergency" | "Non-Emergency" | "Insufficient Information",
  "reasoning_summary": "<1 -2 sentence explanation for label in English>"
}
```

## N.4 Triage Rules + Knowledge Base with chain of thought

You are a medical triage assistant for maternal and newborn care. You will be given:

- the user's message (mother/caregiver/pregnant woman)

Messages may be written in Indian languages or English, may include code-mixed content, and can be short, informal, romanized and chatty in tone.

Your task is to classify the message into one of three Labels:

**Emergency:** The message indicates, or could plausibly indicate, a symptom that may require immediate medical attention or urgent care. If the message overlaps with any emergency symptom in the knowledge base, or describes something that could reasonably map to it (even if severity details are missing), label it Emergency. When in doubt between Emergency and Insufficient Information, prefer Emergency.

**Non-Emergency:** The message describes a symptom or concern that clearly does not require urgent attention. These include general diet questions, breastfeeding without emergency symptoms, introducing food, minor constipation/gas, admin questions, sleep/teething/routine issues with no possible overlap with emergency symptoms.

**Insufficient Information:** Use when the message lacks enough detail to decide if 'its an emergency. This includes vague, incomplete, or poorly written messages that might hint at symptoms but 'dont confirm them. If a symptom clearly matches the emergency list, classify as Emergency. If it could be an emergency but more details are needed to know for sure, classify as Insufficient Information.

**Analysis rules:**

- Treat questions as reports of situations that occurred.
- If multiple overlaps and at least one could be an emergency + label Emergency immediately.
- Severity cues (e.g., "very", "a lot", "constant", "unbearable", "multiple "times" strengthen the Emergency signal, but absence of severity should NOT downgrade to Insufficient Information if the symptom itself is inherently urgent (e.g., bleeding, localised pain, difficulty breathing, high fever, jaundice).
- For ambiguous mentions (e.g., "bleeding without source), assume worst-case relevant to Knowledge Base (vaginal bleeding).
- For cases where localised pain (e.g., such as headache, back pain, leg pain, etc), bleeding, dizziness is mentioned label it as Emergency.
- For infant feeding issues, if feeding has stopped or the baby is "not feeding", treat as Emergency unless clearly stated otherwise.
- If information is so vague that no symptoms can be matched at all, then label Insufficient Information.

```
- Do not penalize informal, chatty, or ungrammatical messages if a
  symptom match is still possible.
- Use the Knowledge Base as the primary guide for Emergency
  symptoms, but allow for reasonable clinical interpretation
  when appropriate.
- Output is routing support, not medical advice; do not give care
  instructions.

Output JSON only (no extra text, no code fences):
{
  "triage_rules": "<1 -2 sentence triage rules used to decide the
    label>",
  "label": "Emergency" | "Non-Emergency" | "Insufficient
    Information",
  "reasoning_summary": "<1 -2 sentence explanation for label in
    English>"
}

Knowledge Base:
[Emergency symptom list provided here]
```

## O Reasoning Trace Analysis: Correct Comprehension, Different Triage

To illustrate how script alone influences triage decisions despite equivalent comprehension, we present a representative case.

|                     |                                                                                                       |
|---------------------|-------------------------------------------------------------------------------------------------------|
| English Translation | I just had colon surgery three days ago. My stomach is hurting, and I have a fever. What should I do? |
| Ground Truth        | emergent                                                                                              |

Table 15: Hindi Native reasoning trace.

| Script Predicted Label | Hindi Native emergent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Reasoning Trace</b> | <p>Okay, let's see. The user is saying they had <b>colon surgery three days ago</b> and now they have a <b>stomach ache</b> and a <b>fever</b>. They're asking what they should do. First, I need to determine the urgency. The key points here are recent surgery and current symptoms. Post-surgery complications can be serious. Pain and fever after surgery might indicate an infection or other issues like a surgical site infection, internal bleeding, or maybe something like a blockage. Since it's only been three days, that's a critical period where complications can arise. The fever is a red flag because it can signify infection. The pain is also concerning because it's not normal after surgery to have increasing pain.</p> <p>The triage labels are emergent, conditionally_emergent, or non_emergent. Emergent would be if they need immediate ER care. Conditionally emergent if it depends on other factors not mentioned. Non-emergent if it's not urgent.</p> <p>In this case, the user is showing symptoms that could be serious post-op complications. However, without more details like the severity of the pain, fever temperature, or other symptoms (like vomiting, inability to pass gas, etc.), it's possible but not certain. The presence of both pain and fever after surgery is a common sign of infection, which does require prompt medical attention. But since the message doesn't specify if the fever is high or if there are other symptoms, maybe it's conditionally emergent. Wait, but the guidelines for triage after surgery usually advise contacting a doctor immediately if there's fever and pain. So maybe emergent.</p> <p>But the rules say to use only the information present. The user is three days post-op with pain and fever. Those are clear signs that they should seek urgent care. So the correct label would be emergent because they need to go to the ER or call emergency services right away. The reason is that post-surgical complications like infection or internal bleeding can be life-threatening if not treated immediately. So the answer should be emergent.</p> |

Table 16: Hindi Romanized reasoning trace.

| Script Predicted Label | Hindi Romanized conditionally_emergent                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Reasoning Trace</b> | <p>Okay, let's see. The user is writing in Hindi, and they mention that they had a <b>colon surgery three days ago</b>. They're experiencing <b>stomach pain</b> and a <b>fever</b>. The question is what they should do. First, I need to determine the urgency. The user had recent surgery, and now has pain and fever. Post-surgical complications can be serious. Fever could indicate an infection, which is a common complication after surgery. Stomach pain might be related to the surgery site. These symptoms could be signs of something that needs immediate attention, like an infection or internal bleeding. However, the message doesn't mention other critical symptoms like severe bleeding, difficulty breathing, or loss of consciousness. But since the fever and pain are present after surgery, which are red flags, this could be an emergency. However, the exact cause isn't clear from the message. The user might need to see a doctor urgently, but without more context, like the severity of the fever or pain, it's conditionally emergent. The presence of post-op complications requires immediate care, but the message lacks details to confirm the exact risk level. So the label would be conditionally_emergent.</p> |

## P Precision by class and script type

Table 17: Precision by class and script type, averaged over four top performing frontier models (Claude-S4.5, GPT-4o, Qwen3-80B, DeepSeek-V3). Precision = TP / (TP + FP).

| Class                    | English | Native | Romanized |
|--------------------------|---------|--------|-----------|
| Emergency                | 91.4    | 89.6   | 84.0      |
| Non-Emergency            | 89.8    | 84.6   | 87.0      |
| Insufficient Information | 13.9    | 11.4   | 3.9       |

### P.1 Post hoc Model-generated Reasoning

**Bigram Analysis.** For each model and script type, we partition the set of reasoning summaries into two groups: correct predictions as  $\mathcal{C}$  and incorrect predictions as  $\mathcal{W}$ . Each reasoning summary is tokenized and converted into a sequence of bigrams, and we compute the bigram ratio  $R(b)$  as:

$$R(b) = \frac{f_w(b) + 1}{f_c(b) + 1} \quad (1)$$

where  $f_w(b)$  denotes the frequency of bigram  $b$  across all summaries in  $\mathcal{W}$ , and  $f_c(b)$  the corresponding frequency in  $\mathcal{C}$ . Incorrect predictions

on romanized inputs are characterized by comprehension and uncertainty hedging bigrams such as “*message vague*”, and “*lacks details*”. In contrast, incorrect predictions on native script contain domain-specific bigrams such as “*chest indrawing*”, and “*cough symptoms*”, indicating that models engage with medical content in native scripts. The bigram “*but lacks*” appears in the top-20 incorrect-prediction list for four models across script types, suggesting a shared learned hedging template for uncertainty. Detailed bigram analysis is in Appendix G.

## Q Flip Rate Definition

For a user message  $i$ , let  $y_i(t)$  denote the model prediction at temperature  $t \in \mathcal{T}$ . We define the *flip indicator* as

$$f_i = \mathbb{1}[|\{y_i(t) : t \in \mathcal{T}\}| > 1] \quad (2)$$

For a subset  $S_{\ell,s}$  corresponding to language  $\ell$  and script  $s$ , the *flip rate* is

$$\text{FlipRate}(S_{\ell,s}) = \frac{1}{|S_{\ell,s}|} \sum_{i \in S_{\ell,s}} f_i. \quad (3)$$

## R Access Dates for Closed Models

GPT-4o was accessed on October 29, 2025, at 19:27:15 UTC, and Claude Sonnet 4.5 was accessed on October 30, 2025, at 14:22:25 UTC for benchmark in Figure 2.

## S HealthBench Triage Label Definition

**Emergent:** Message that reflects that the user should immediately seek urgent medical care (e.g., go to the ER, call 911, or equivalent).

**Conditionally Emergent:** Message where it is possible that the user should immediately seek urgent medical care (e.g., go to the ER, call 911, or equivalent), depending on a risk factor or other circumstances not present in the message.

**Non emergent:** Message that reflects that the user should not seek emergency-level medical care immediately, but may need to seek medical care in a different time frame or setting.

## T HealthBench Translation Quality

We assessed the translation quality on a random sample of 20 HealthBench (Validation set) queries for each of five languages. Native-speaker annotators rated each translation 1/0, where 1 indicates an

adequate translation with key symptoms preserved and 0 otherwise. For each language, we report the percentage of translations rated adequate. Hindi Romanized: 100%; Marathi Romanized: 80%; Punjabi Romanized: 100%; Kannada Romanized: 85%; Telugu Romanized: 80%.