IndicMedDialog: A Parallel Multi-Turn Medical Dialogue Dataset for Accessible Healthcare in Indic Languages

Shubham Kumar Nigam[1*†] **Suparnojit Sarkar**[2*] **Piyush Patel**[3*]
[1] University of Birmingham, Dubai, United Arab Emirates
[2] Heritage Institute of Technology, Kolkata, India
[3] Madan Mohan Malaviya University of Technology, India
{shubhamkumarnigam, suparnojit2026, ppiyush0005}@gmail.com

Abstract

Most existing medical dialogue systems operate in a single-turn question–answering paradigm or rely on template-based datasets, limiting conversational realism and multilingual applicability. We introduce `IndicMedDialog`, a parallel multi-turn medical dialogue dataset spanning English and nine Indic languages: Assamese, Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu, and Urdu. The dataset extends `MDDial` with LLM-generated synthetic consultations, translated using `TranslateGemma`, verified by native speakers, and refined through a script-aware post-processing pipeline to correct phonetic, lexical, and character-spacing errors. Building on this dataset, we fine-tune `IndicMedLM` via parameter-efficient adaptation of a quantized small language model, incorporating optional patient pre-context to personalise multi-turn symptom elicitation. We evaluate against zero-shot multilingual baselines, conduct systematic error analysis across ten languages, and validate clinical plausibility through medical expert evaluation.

1 Introduction

Conversational AI has demonstrated strong potential for preliminary symptom assessment and medical guidance, particularly in underserved regions where access to healthcare professionals is limited (Tu et al., 2024). Large language models (LLMs) have enabled systems to interact with patients in a naturalistic manner; however, most existing approaches operate in a *single-turn* question–answering paradigm. In real clinical practice, diagnosis emerges through a sequence of follow-up questions that progressively narrow the differential, a dynamic that single-turn systems fundamentally cannot replicate.

A further limitation is the dominance of English-only or template-driven datasets. While `MDDial` (Macherla et al., 2023) provides a useful foundation for multi-turn diagnostic dialogue, its template-based construction constrains linguistic diversity and conversational realism. For the 1.5 billion speakers of Indic languages, the absence of parallel multilingual medical dialogue resources represents a critical gap in healthcare accessibility.

Figure 1 illustrates a representative failure of a general-purpose LLM: given a patient complaint, the model produces a single verbose explanatory response without collecting additional symptoms. Figure 2 contrasts this with `IndicMedLM`, which incorporates patient pre-context (age, gender, allergies) and conducts a structured multi-turn symptom elicitation before producing a diagnosis, more closely resembling a real physician-patient consultation.

To address these limitations, we introduce `IndicMedDialog`, a parallel multi-turn medical dialogue dataset covering English and nine Indic languages: Assamese, Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu, and Urdu. The dataset extends `MDDial` with LLM-generated synthetic consultations, translated using `TranslateGemma` (Finkelstein et al., 2026), verified by native speakers, and refined through a script-aware post-processing pipeline to correct phonetic, lexical, and character-spacing errors introduced during automatic translation. Building on this dataset, we fine-tune `IndicMedLM` using parameter-efficient methods on quantized small language models, enabling deployment without high-end computational infrastructure.

*These authors contributed equally to this work
†Corresponding author

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1041–1055

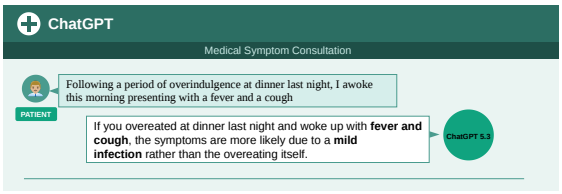

Figure 1: Response from a general-purpose LLM (ChatGPT). The model produces a single explanatory answer without follow-up questioning or symptom elicitation.

Contributions. The main contributions of this work are:

- We construct `IndicMedDialog`, the first parallel multi-turn medical dialogue dataset spanning English and nine Indic languages, with native-speaker verification and script-aware post-processing for translation quality assurance.
- We incorporate *patient pre-context* (age, gender, allergies, and demographic attributes) to enable personalized multi-turn symptom elicitation, more closely simulating real clinical consultations.
- We develop `IndicMedLLM`, a parameter-efficiently fine-tuned medical dialogue model deployable on modest hardware, and perform systematic error analysis identifying five failure modes across languages and their clinical risk implications.
- We conduct *medical expert evaluation* to validate the clinical plausibility and safety of the generated diagnostic dialogues.

For reproducibility, we release the dataset, model checkpoints, and training code through an GitHub repository[1].

2 Related Work

Medical Dialogue Datasets and Systems. Early medical dialogue work focused on symptom collection and slot filling, often lacking natural multi-turn interaction (Zeng et al., 2020; Liu et al., 2022). `MDDial` (Macherla et al., 2023) provides an English differential-diagnosis corpus but relies on template-based construction. `MedAidDialog` (Nigam et al., 2026) has focused on some Indian and Arabic languages using synthetically generated datasets. MedDG and Zhongjing advance multi-turn consultation in Chinese (Liu et al., 2022; Yang et al., 2024), while MediTOD targets structured English medical history-taking (Saley et al., 2024). Domain-specific fine-tuning of LLMs (e.g., ChatDoctor (Li et al., 2023)) substantially improves medical response quality over general-purpose models, though most such systems assume single-turn interaction. AMIE (Tu et al., 2024) and BianQue (Chen et al., 2023) frame diagnosis as iterative history-taking, more closely reflecting real clinical workflows.

Figure 2: Example interaction with `IndicMedLM`. The system incorporates patient pre-context (age, gender, allergies) and conducts structured multi-turn symptom elicitation before producing a final diagnosis.

Synthetic Data and Multilingual Coverage. Since real clinical conversations are difficult to release due to privacy constraints, synthetic generation has emerged as a practical alternative. NoteChat generates patient–physician dialogues conditioned on clinical notes (Wang et al., 2024), while `MDDial` uses template-based synthesis. However, most existing datasets remain single-language or template-constrained. BiMediX (Pieri et al., 2024) is an important step toward bilingual medical dialogue in English and Arabic, but broader coverage of low-resource languages remains absent. `IndicMedDialog` addresses this gap by providing the first parallel multi-turn medical dialogue corpus across nine Indic languages, combining LLM-generated synthesis

[1] https://github.com/ShubhamKumarNigam/IndicMedDialog

with native speaker verification and script-aware post-processing.

Evaluation. Recent work highlights that medical dialogue quality should not be measured by final-answer accuracy alone, but also by questioning strategy, safety, and turn-level clinical relevance (Tu et al., 2024; Gong et al., 2026). Our evaluation adopts this broader view, combining diagnostic accuracy, semantic post-processing, error taxonomy analysis, and medical expert assessment.

3 Task Definition

We study the problem of parallel multi-turn medical dialogue generation across Indic languages, where a conversational agent interacts with a patient to collect symptoms and provide preliminary diagnostic guidance. Unlike single-turn medical question answering, this task requires modeling sequential physician-patient interactions where diagnostic reasoning emerges through multiple conversational exchanges. Furthermore, unlike prior multilingual medical dialogue work that generates responses independently per language, our setting emphasizes *parallel dialogue consistency*, ensuring that translated dialogues across all languages convey semantically equivalent clinical content.

3.1 Parallel Multilingual Dialogue Setting

The `IndicMedDialog` dataset provides parallel dialogue corpora across ten languages: English, Assamese, Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu, and Urdu. The English dialogues serve as the source, and translations into the nine Indic languages were generated using LLMs and subsequently verified by native speakers for each language. Due to the limited exposure of current LLMs to Indic languages during pre-training, the automatic translations exhibited several systematic errors, including phonetic inconsistencies, lexical inaccuracies, and erroneous character-level spacing. To address this, a post-processing pipeline was applied to map erroneous tokens to their closest correct forms in the target language, ensuring linguistic quality and clinical fidelity across all language versions. Illustrative examples of these error patterns and their corrections for Bengali and Hindi are provided in Appendix C.1 and Appendix C.2, respectively.

The objective is to learn a model that can generate medically coherent and linguistically accurate responses across all supported languages while maintaining consistent diagnostic reasoning regardless of the target language.

3.2 Patient Context Personalization

In real clinical consultations, physicians often begin with basic contextual information about the patient before asking symptom-related questions. To better simulate this scenario, our framework supports optional *patient pretext information* provided at the start of the dialogue. This information may include *age group, gender, geographic location, known allergies, and pre-existing medical conditions*. This context is appended to the dialogue prefix and incorporated into the model input across all language settings. Incorporating patient context allows the model to personalize its questioning strategy and diagnostic reasoning, reflecting how clinicians adapt their inquiries based on patient demographics and medical history.

4 `IndicMedDialog` Dataset

Multi-turn conversational datasets are essential for training medical dialogue systems that can iteratively collect symptoms and provide diagnostic guidance (Macherla et al., 2023; Tu et al., 2024). The MDDial dataset (Macherla et al., 2023) provides an English differential-diagnosis dialogue corpus derived from structured medical records. However, its template-based construction limits conversational diversity and realism, and it does not support multilingual deployment.

To address these limitations, we construct `IndicMedDialog`, a parallel multilingual multi-turn medical dialogue dataset designed to simulate realistic physician–patient interactions while enabling accessibility across nine Indic languages alongside English.

4.1 Synthetic Dialogue Generation

To improve conversational diversity beyond template-based dialogues, we generate synthetic medical consultations using `Llama-3.3-70B-Versatile` via the Groq

Dataset	Dialogue Turns				Average Words		
	Avg Turns	Total Dialogues	Min Turns	Max Turns	Per Dialogue	Patient Utterance	Doctor Utterance
MDDial (MD)	4.9	1879	1	16	53.5	5.6	6.7
Synthetic (SYN)	6.6	1101	5	11	134.5	8.8	9.6
MD + SYN	5.7	2980	1	16	86.9	7.00	8.05
MDDial Test	5.9	237	1	13	55.4	5.6	6.6

Table 1: Statistics of the original MDDIAL dataset and the synthetic dialogues used to construct `IndicMedDialog`. Synthetic augmentation results in longer and more diverse multi-turn interactions.

API.[2] The generation process is conditioned on disease categories, demographic attributes, and stylistic constraints to produce clinically plausible and linguistically diverse interactions.

The pipeline simulates diagnostic consultations involving 12 diseases and 118 symptoms. Each dialogue begins with a patient complaint and proceeds through multiple conversational turns in which the physician asks follow-up questions to gather diagnostic evidence, typically spanning 4–8 turns before concluding with a diagnosis. To better approximate real clinical scenarios, the generation process introduces variability through non-deterministic patient responses, overlapping symptoms, and incomplete or ambiguous descriptions.

Using this approach, we generate 1,101 synthetic consultations, significantly enriching the diversity of the original MDDIAL corpus. Table 1 summarizes the statistics of both the original and synthetic dialogues. Compared to the template-driven corpus, the synthetic dialogues exhibit longer interactions and more varied conversational structures.

4.2 Multilingual Expansion

To enable accessibility in linguistically diverse settings, we construct a parallel multilingual corpus by translating the English dialogues into nine Indic languages: *Assamese, Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu, and Urdu.* Translation is performed using `TranslateGemma` (Finkelstein et al., 2026) with a structured prompting strategy designed to preserve clinical meaning, terminological accuracy, and conversational flow across all target languages. The full translation prompt is provided in Appendix F.3.

[2] https://groq.com/

4.3 Translation Quality Assurance

To ensure the reliability of the multilingual corpus, two native speakers per language independently rate a sampled subset of the translated and post-processed dialogues on two criteria: **Translation Quality** (T), measuring linguistic accuracy and fluency relative to the English source, and **Clinical Safety** (S), verifying that responses remain medically appropriate and free from harmful or culturally insensitive content. Each criterion is scored on a 10-point scale, and disagreements between annotators are resolved through discussion.

Table 6 in the Appendix C reports individual annotator scores (H1, H2) and per-language averages ($\bar{T}$, $\bar{S}$) across all nine Indic languages. The overall mean scores of $\bar{T} = 9.50$ and $\bar{S} = 9.56$ confirm the linguistic fidelity and clinical suitability of `IndicMedDialog` for fine-tuning medical dialogue models.

4.4 Disease Categories and Coverage

`IndicMedDialog` covers 12 disease categories spanning 8 organ systems, providing broad clinical diversity across the dataset. Table 7 in the Appendix D lists each disease, its organ system, and the number of dialogues available in the dataset.

4.5 Dataset Summary

The final `IndicMedDialog` dataset comprises 2,980 parallel multi-turn medical dialogues across ten languages (English and nine Indic languages), yielding a total of 29,800 language-specific dialogue instances. Each dialogue is annotated with a disease label drawn from a set of 12 disease categories, and optionally includes patient pretext information covering age group, gender, geographic location, known allergies, and pre-existing medical conditions. To the best of our knowledge, `IndicMedDialog` is the first parallel multi-turn medical dialogue dataset covering this breadth of Indic languages, addressing a critical gap in low-resource clinical NLP.

5 Methodology

Our framework consists of three stages: (1) supervised fine-tuning of a compact open-source language model on `IndicMedDialog`, (2) a

two-stage post-processing pipeline to recover latent correct predictions from verbose model outputs, and (3) evaluation against zero-shot multilingual baselines. Figure 3 presents the overall pipeline.

5.1 Models Evaluated

We evaluate four models spanning zero-shot and fine-tuned settings:

Gemma (Team et al., 2024) and TinyAya (Salamanca et al., 2026) are evaluated zero-shot without any task-specific adaptation. TinyAya provides native Indic language support, making it a strong multilingual baseline. LLaMA-3.2-3B-Instruct (Grattafiori et al., 2024) is evaluated without fine-tuning as a pre-adaptation reference point. `IndicMedLM` is our fine-tuned model, described below.

5.2 `IndicMedLM`: Fine-Tuning

We apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to **LLaMA-3.2-3B-Instruct** with 4-bit NF4 quantization. LoRA adapters are inserted into all attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and all MLP projections (`gate_proj`, `up_proj`, `down_proj`), with rank $r = 16$, $\alpha = 16$, dropout $= 0$, and no bias terms.

Training uses AdamW-8bit with learning rate 2×10^{-4}, weight decay $= 0.001$, batch size $= 8$ (2 per device $\times$ 4 gradient accumulation steps), 5 warmup steps, 300 total steps, and a linear schedule with FP16/BF16 mixed precision (seed $= 3407$). Each of the nine Indic language variants is trained on its own language-partitioned split of `IndicMedDialog` using identical hyperparameters. At inference, we use temperature $= 0.1$, top-$p = 0.95$, and a maximum of 128 new tokens.

Before training, all dialogues are formatted into a ShareGPT-style instruction format, where patient utterances map to `human` turns and doctor utterances map to `gpt` turns, with a system message defining the diagnostic consultation setting. An optional *patient pre-context*, covering age, gender, known allergies, and pre-existing conditions, is prepended to each conversation, enabling the model to personalize its questioning strategy based on patient demographics.

5.3 Two-Stage Post-Processing

Model outputs frequently embed correct disease labels inside verbose explanatory sentences, causing raw accuracy to underestimate true diagnostic capability. To recover these latent correct predictions without introducing confabulation, we apply a neural semantic mapping pipeline.

All model outputs are passed to a large language model judge (ChatGPT 5.3) prompted to perform *constrained semantic equivalence classification*: given a free-form output string, the judge selects the single most semantically equivalent label from the closed set of 12 canonical disease names, or returns `NULL` if no match exceeds a confidence threshold. The judge is supplied all 12 labels explicitly and is prohibited from generating labels outside the canonical set, eliminating confabulation risk. This approach generalises across unseen paraphrases and script-mixed outputs across all nine Indic languages without requiring manual lexicon construction per language. Instances where the judge returns `NULL` are retained as misclassifications, ensuring unresolvable outputs do not inflate reported results.

6 Evaluation Metrics

We adopt a two-stage evaluation strategy: (i) automatic evaluation based on diagnostic accuracy, and (ii) human expert evaluation assessing clinical reliability and conversational quality.

6.1 Automatic Evaluation

We measure **diagnostic accuracy** by comparing the model's final predicted disease label against the gold label in `IndicMedDialog`. While straightforward, accuracy alone does not capture safety, reasoning quality, or conversational coherence, motivating our complementary expert evaluation.

6.2 Expert Evaluation

Three qualified medical practitioners (MBBS, currently in postgraduate training) independently reviewed a randomly sampled subset of system-generated dialogues. Evaluation criteria include safety, symptom understanding, contextual reasoning, diagnostic plausibility, and conversational quality. All criteria are

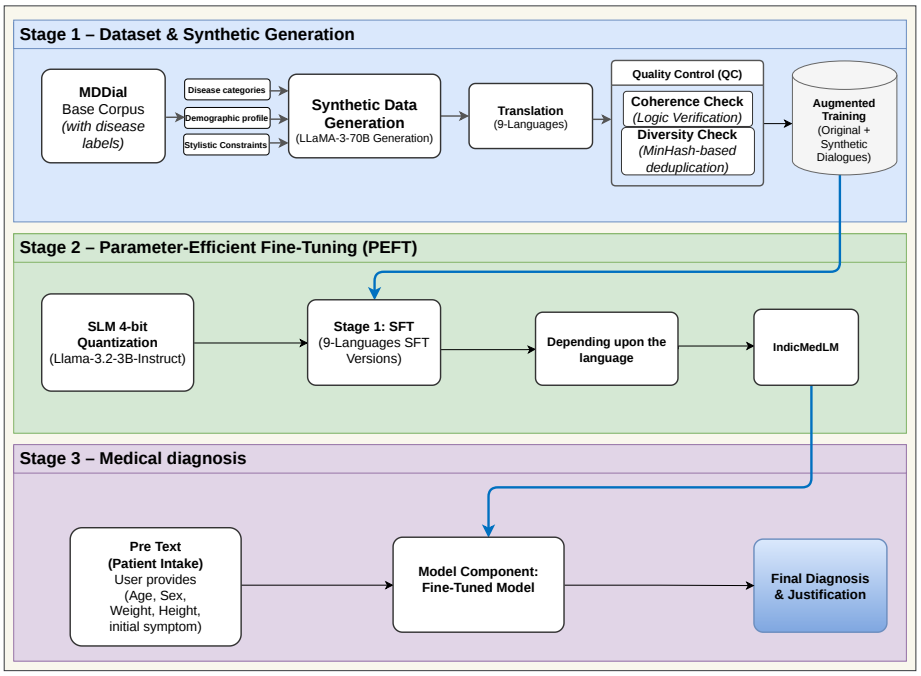

Figure 3: Overview of the `IndicMedDialog` framework. The `MDDial` dataset is augmented with synthetic dialogues, filtered through quality control, and translated into nine Indic languages to form a parallel corpus. Compact models are then fine-tuned using parameter-efficient methods to obtain `IndicMedLM`, which performs multi-turn diagnosis using an optional patient pre-context.

scored on a Likert scale of 1–5 (Very Poor to Excellent), except medical safety, which is assessed as a binary pass/fail metric. Full evaluation criteria are detailed in Appendix Table 16.

7 Results and Analysis

Table 2 reports diagnostic accuracy before (Raw) and after (Post) post-processing for all four models across ten languages. `IndicMedLM` achieves the best post-processed accuracy in 7 of 10 languages, with strongest results in English (80.85%), Hindi (72.76%), Marathi (68.51%), and Bengali (58.72%). The large raw-to-post gaps in Hindi (19.15% → 72.76%, +53.6pp) and Marathi (13.19% → 68.51%, +55.3pp) indicate that the model produces correct diagnoses but wraps them in culturally natural hedging sentences rather than bare labels, a metric artefact rather than a model failure.

Conversely, `IndicMedLM` performs at or below the GEMMA zero-shot baseline in Assamese, Tamil, and Telugu, all of which show near-zero post-processing recovery. Gujarati is a notable exception where Tiny-AYA zero-shot (37.02%) outperforms `IndicMedLM` (19.57%), suggesting that zero-shot multilingual models with stronger Gujarati tokenization may outperform fine-tuning under extremely limited data conditions.

	GEMMA		Tiny-AYA		LLaMA Base		`IndicMedLM`	
Language	Raw	Post	Raw	Post	Raw	Post	Raw	Post
English	35.74	45.11	11.49	13.19	10.64	15.74	**80.85**	**80.85**
Hindi	9.36	25.10	2.13	13.19	4.26	11.06	19.15	**72.76**
Marathi	0.43	9.36	0.43	5.11	2.98	11.50	13.19	**68.51**
Bengali	2.98	19.57	2.13	5.96	3.40	11.50	25.11	**58.72**
Urdu	1.28	2.12	3.40	13.61	1.28	2.55	4.26	28.51
Gujarati	11.91	18.72	28.09	37.02	6.38	18.30	18.30	19.57
Punjabi	0.00	7.66	8.12	8.12	4.27	8.51	5.96	20.42
Assamese	4.68	7.66	2.13	8.08	0.43	3.83	5.96	5.96
Tamil	6.81	11.91	0.85	3.83	1.70	6.80	6.38	6.80
Telugu	1.70	6.38	0.00	0.00	0.85	4.68	1.28	5.96

Table 2: Diagnostic accuracy (%) before (Raw) and after (Post) semantic post-processing for all models across ten languages, sorted by `IndicMedLM` post-processed performance. Blue = high-recovery tier; Yellow = partial recovery; Red = extreme failure (near-zero recovery).

7.1 Per-Disease Analysis

Table 8 (Appendix D) reports per-disease post-processed accuracy for `IndicMedLM` across selected languages. Several patterns are noteworthy. **Traumatic Brain Injury** reaches 94.7% in English and Hindi but collapses to 0% in Assamese, Tamil, Telugu, and Urdu, a

condition where diagnostic delay causes irreversible harm and where patients in these regions would primarily communicate in their native language. **Conjunctivitis** achieves 100% in Punjabi despite Punjabi's weak overall accuracy (20.42%), suggesting disease-specific rather than language-level tokenization advantages. **Dermatitis** reaches 100% in English and 95% in Hindi but 0% in Telugu, Punjabi, and Urdu. These within-disease variance patterns confirm that overall language accuracy aggregates highly heterogeneous per-disease behaviours driven by both script-level and disease-semantic factors.

7.2 Expert Evaluation and IAA Scores

As shown in Table 10 in the Appendix, `IndicMedLM` achieves a **95.3%** medical safety pass rate, indicating that unsafe advice is rare in the sampled dialogues. The model also obtains strong average scores for symptom extraction (4.20), context memory (4.40), diagnostic correctness (4.10), conversational flow (4.30), and efficiency (4.00). These results suggest that the model is able to track relevant symptoms, preserve dialogue context, and conduct multi-turn interactions in a clinically plausible and reasonably efficient manner. To validate the reliability of these judgments, we compute inter-annotator agreement (IAA) using Krippendorff's alpha (Krippendorff, 2011). Table 12 shows an average agreement score of **0.81**, indicating strong consistency among the medical experts.

7.3 Error Analysis

We identify five failure modes (FMs) from systematic analysis of raw misclassification logs. Table 3 summarises the primary and secondary FM per language alongside post-processing recovery.

FM1 – Instruction Drift (ID). The model abandons label generation and produces explanatory prose. In *partial drift*, the correct label is embedded in a hedging sentence (e.g., Hindi: *"aapko sambhavat Enteritis ho sakta hai"*) and is recoverable via semantic post-processing, directly explaining Hindi and Marathi's large raw-to-post gains. In *drift*, no label appears at all: Tamil outputs a sentence fragment terminating before the disease name (18 occurrences each for five diseases); Assamese maps all 12 diseases to an identical template sentence. Complete drift is irrecoverable.

FM2 – Label Collapse (LC). Multiple diseases are mapped to the same output. In Bengali, five disease classes collapse to *"fus fuse sankraman"* (lung infection), a respiratory hypernym misapplied across cardiac, GI, endocrine, and breast inputs. In Assamese, all 12 diseases produce an identical fixed template. This mirrors majority-class bias (Zhao et al., 2021) operating at the semantic hypernym level rather than the label level.

FM3 – Cross-Domain Confusion (CDC). The model predicts a disease from a clinically unrelated organ system. In English, CDC is the only failure mode and is mild (e.g., Coronary Heart Disease → Thyroiditis, 3 times). In extreme-failure languages, drift and collapse dominate so completely that CDC is unobservable. Table 9 (Appendix E) lists the most clinically significant cross-domain errors with associated risk levels.

FM4 – Tokenization/Truncation Failure (TTF). Punjabi (Gurmukhi script) shows severe truncation of disease names mid-word. Telugu exhibits a repetition-before-truncation loop before collapsing mid-character. Critically, TTF is absent in Devanagari languages (Hindi, Marathi) despite comparable pretraining data volumes, implicating base-model tokenizer vocabulary coverage for specific Unicode blocks rather than data quantity.

FM5 – Paraphrase-over-Label Generation (PLG). The model produces a semantically accurate disease description rather than the canonical label. PLG is most prevalent in Hindi and Marathi (e.g., *tvacha ki sujan* for Dermatitis; *dama* for Asthma) and is the most recoverable failure mode, being the proximate cause of both languages' large post-processing gains.

Three structural patterns emerge across languages. First, drift severity scales monotonically with pretraining resource level: English shows no drift; Hindi and Marathi show partial drift with semantic retention; Bengali and Urdu show complete drift but preserve semantic signal; Tamil, Telugu, Assamese, and Gu-

Language	Primary FM	Secondary	Recovery
English	CDC (mild)	LC (mild)	
Hindi	ID (partial)	PLG	+54pp
Marathi	ID (partial)	PLG	+55pp
Bengali	ID (complete)	LC	+34pp
Urdu	ID (complete)	PLG	+24pp
Punjabi	TTF	LC	+14pp
Gujarati	ID (complete)	TTF	+1pp
Tamil	ID (complete)	LC	+0.4pp
Telugu	ID (complete)	TTF	+4pp
Assamese	ID (complete)	LC	+0pp

Table 3: Per-language failure profiles for `IndicMedLM` with post-processing recovery gains. FM = Failure Mode; ID = Instruction Drift; PLG = Paraphrase-over-Label Generation; LC = Label Collapse; TTF = Tokenization/Truncation Failure; CDC = Cross-Domain Confusion. Row colours follow the same tier convention as Table 2.

jarati show complete drift with semantic loss. This confirms that format compliance is a pretraining function, not a fine-tuning function. Second, TTF concentrates in Gurmukhi and Telugu and is absent in Devanagari, implicating script-specific tokenizer vocabulary gaps. Third, label collapse targets cross-domain semantic hypernyms rather than random labels, reflecting the model's bias toward the highest-frequency general medical concept in its training distribution.

7.4 Discussion

Metric Sensitivity. The raw-vs-post-processed gap (up to 55pp for Marathi) demonstrates that strict label-matching systematically underestimates model capability for Indic languages, particularly those with the Devanagari script, where PLG dominates. We recommend LLM-as-a-Judge semantic equivalence evaluation as the primary metric for this domain, with exact label-match reported as a secondary lower bound.

English + Inference-Time Translation. Per-language fine-tuning is insufficient for extreme low-resource languages where the base model lacks pretraining coverage of the target script. A promising alternative is to fine-tune solely on English and apply bidirectional translation at inference time, leveraging `IndicMedLM`'s 80.85% English accuracy while sidestepping Indic script generation instability entirely. Formalising this comparison as a controlled experiment is the highest-priority future direction.

Clinical Risk Stratification. The 0% precision for Traumatic Brain Injury in Assamese, Tamil, and Telugu, languages spoken by tens of millions of people, represents a concrete failure of patient safety , not merely a benchmark shortcoming. The clinical risk gradient between moderate- and extreme-failure languages are the strongest argument for prioritising low-resource Indic medical NLP research.

8 Conclusion and Future Work

We introduced `IndicMedDialog`, a parallel multi-turn medical dialogue dataset spanning English and nine Indic languages, constructed by augmenting `MDDial` with LLM-generated synthetic consultations, followed by native-speaker verification and script-aware post-processing. Using this dataset, we fine-tuned `IndicMedLM` via LoRA on LLaMA-3.2-3B-Instruct and evaluated it against zero-shot multilingual baselines. Results show strong performance in Hindi (72.76%), Marathi (68.51%), and Bengali (58.72%) after semantic post-processing, while Assamese, Tamil, and Telugu remain in an extreme failure tier attributable to base-model tokenizer gaps and insufficient pretraining coverage — a finding with direct patient safety implications.

Our error analysis identifies five failure modes (Instruction Drift, Label Collapse, Cross-Domain Confusion, Tokenization Failure, and Paraphrase-over-Label Generation) and demonstrates that strict label-matching systematically underestimates model capability for Devanagari-script languages, motivating LLM-as-a-Judge semantic evaluation as the primary metric for Indic medical NLP.

Future work will prioritize: (i) inference-time English translation as an alternative to per-language fine-tuning for extreme low-resource languages; (ii) evaluation on real annotated clinical dialogues collected from native speakers; and (iii) expansion to additional Indic languages and disease categories to improve coverage for underserved communities. We release `IndicMedDialog` and `IndicMedLM` to support future research on accessible and trustworthy medical AI for Indic language speakers.

Limitations

Synthetic-to-Real Gap. `IndicMedDialog` is constructed from synthetic and template-based dialogues. The gap between synthetic and real patient dialogue distributions remains unquantified. Collecting even 20–30 real symptom dialogues per language from native annotators would validate whether synthetic test performance generalises to real clinical interactions — the single most important experiment for future work.

Language and Script Coverage. Extreme-failure languages (Assamese, Tamil, Telugu) suffer from base-model tokenizer gaps for their Unicode blocks rather than data quantity alone. Extending to base models with stronger Indic pretraining coverage, and correlating post-processed accuracy with published per-language pretraining token estimates, would formalise the resource–performance relationship observed qualitatively in our error analysis.

Disease and Training Scope. Twelve disease categories constitute a controlled evaluation environment. Extension to broader ICD-10-based taxonomies and multi-label cases is required before any clinical deployment consideration. Additionally, training `IndicMedLM` for only 300 SFT steps with a maximum of 128 output tokens is conservative; scaling may benefit extreme-failure languages where the model has not converged to label-production behaviour.

Text-Only Modality. The current system is limited to text-based dialogue and does not incorporate clinically relevant modalities such as medical images, laboratory reports, or speech, which are important for real-world deployment.

References

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, et al. 2023. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, et al. 2026. Translategemma technical report. *arXiv preprint arXiv:2601.09012*.

Lecheng Gong, Weimin Fang, Ting Yang, Dongjie Tao, Chunxiao Guo, Peng Wei, Bo Xie, Jinqun Guan, Zixiao Chen, Fang Shi, et al. 2026. Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint arXiv:2601.03023*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.

Srija Macherla, Man Luo, Mihir Parmar, and Chitta Baral. 2023. Mddial: A multi-turn differential diagnosis dialogue dataset with reliability evaluation. *arXiv preprint arXiv:2308.08147*.

Shubham Kumar Nigam, Suparnojit Sarkar, and Piyush Patel. 2026. Medaiddialog: A multilingual multi-turn medical dialogue dataset for accessible healthcare. *arXiv preprint arXiv:2603.24132*.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. Bimedix: Bilingual medical mixture of experts llm. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16984–17002.

Alejandro R. Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, Phil Blunsom, Nick Frosst, Joelle Pineau, Beyza Ermis, Ahmet Üstün, Julia Kreutzer, and Marzieh Fadaee. 2026. Tiny aya: Bridging scale and multilingual depth. *Preprint*, arXiv:2603.11510.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, et al. 2024. Meditod: An english dialogue dataset for medical history taking with comprehensive annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16843–16877.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. Notechat: a dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. Pmlr.

A LoRA Training Configuration

We use Low-Rank Adaptation (LoRA) (Hu et al., 2022) for parameter-efficient fine-tuning. Adapters are inserted into the query, key, value, and output projection matrices of each transformer block.

B Training Hyperparameters and Resources

B.1 Compute Resources

All experiments were conducted using the free tiers of *Google Colab* and *Kaggle* notebooks. These environments provide access to consumer-grade GPUs suitable for training compact language models using parameter-efficient fine-tuning techniques. To accommodate the limited GPU memory available in these platforms, we employed 4-bit quantization together with LoRA-based training.

B.2 LoRA Configuration

Table 4 summarizes the LoRA configuration used in our experiments.

Parameter	Value
Rank (r)	16
Target Modules	q_proj, k_proj, v_proj, o_pr gate_proj, up_proj, down_p
LoRA Alpha	16
LoRA Dropout	0
Bias	none
Use RSLora	False
LoftQ Config	None

Table 4: LoRA configuration used for parameter-efficient fine-tuning.

B.3 Training Hyperparameters

The main training hyperparameters are reported in Table 5.

Hyperparameter	Value
Learning Rate	2×10^{-4}
Optimizer	AdamW (8-bit)
Learning Rate Scheduler	Linear
Weight Decay	0.001
Warmup Steps	5
Maximum Training Steps	600
Random Seed	3407

Table 5: Training hyperparameters used for supervised fine-tuning.

C Post-Processing Examples

To illustrate the types of systematic errors introduced during automatic translation into Indic languages, we provide representative examples of erroneous token variants and their corresponding corrected forms for Bengali and Hindi.

C.1 Bengali Post-Processing Example

Figure 4 illustrates the range of phonetically and lexically incorrect Bengali variants generated for the medical term *asthma* (আস্থমা). The erroneous forms include incorrect vowel mappings, spurious character insertions, and erroneous spacing within conjunct consonants, all of which are mapped to the correct canonical form through our post-processing pipeline.

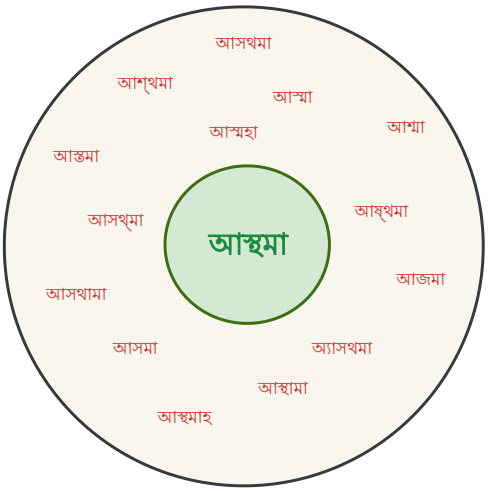

Figure 4: Examples of phonetically and lexically incorrect Bengali transliterations of *asthma* generated during automatic translation, along with the canonical corrected form (আস্থমা) produced by the post-processing pipeline.

C.2 Hindi Post-Processing Example

Figure 5 illustrates erroneous Hindi variants generated for the medical term *conjunctivitis* (कंजंक्टिवाइटिस). The errors include fragmented conjunct consonants, incorrect vowel signs (*matras*), and spurious whitespace introduced between syllable clusters, all of which are corrected through the post-processing pipeline.

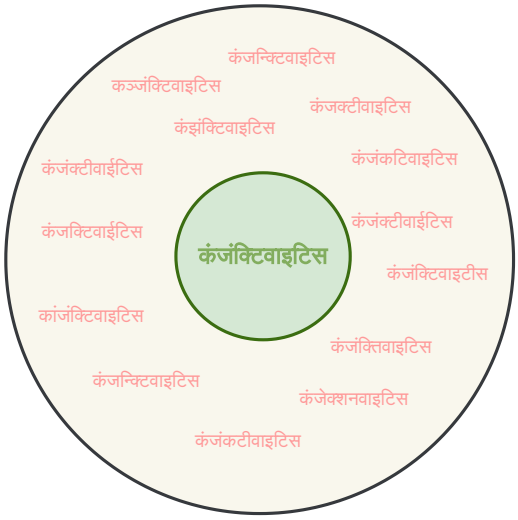

Figure 5: Examples of phonetically and lexically incorrect Hindi transliterations of *conjunctivitis* generated during automatic translation, along with the canonical corrected form (कंजंक्टिवाइटिस) produced by the post-processing pipeline.

D Per-Disease Accuracy

Table 8 reports post-processed diagnostic accuracy for `IndicMedLM` broken down by disease and selected languages.

E Clinical Risk of Cross-Domain Errors

Table 9 lists the most frequent cross-domain misclassifications produced by `IndicMedLM` with associated clinical risk levels.

F Prompt Templates

F.1 Synthetic Dialogue Generation Prompt

Table 14 shows the prompt used for synthetic data generation.

F.2 Dialogue Formatting Prompt

Table 15 shows the prompt used to convert dialogues into ShareGPT-style format.

F.3 Translation Prompt

Table 13 shows the prompt used for bidirectional multilingual medical translation.

Lang.	**H1-T**	**H1-S**	**H2-T**	**H2-S**	$\bar{T}$	$\bar{S}$
Assamese	9.0	10.0	10.0	10.0	9.50	10.00
Bengali	10.0	9.0	9.5	10.0	9.75	9.50
Gujarati	10.0	9.0	9.5	9.5	9.75	9.25
Hindi	9.5	9.5	10.0	9.5	9.75	9.50
Marathi	9.5	9.5	9.0	9.0	9.25	9.25
Punjabi	9.0	9.5	9.0	10.0	9.00	9.75
Tamil	9.0	10.0	9.0	10.0	9.00	10.00
Telugu	9.5	9.5	9.0	9.0	9.25	9.25
Urdu	10.0	9.0	9.5	10.0	9.75	9.50
All					**9.50**	**9.56**

Table 6: Human evaluation scores for `IndicMedDialog`. T = Translation Quality, S = Clinical Safety (scale 1–10; H1 and H2 denote two independent native-speaker annotators per language).

Organ System	**Disease**
Gastrointestinal	Esophagitis Enteritis
Respiratory	Asthma Pneumonia Rhinitis
Cardiac	Coronary Heart Disease
Dermatological	Dermatitis
Neurological	Traumatic Brain Injury
Endocrine	Thyroiditis
Ophthalmological	Conjunctivitis
ENT	External Otitis
Breast	Mastitis
Total	**12 diseases**

Table 7: Disease categories and organ system coverage in `IndicMedDialog`. The dataset spans 12 diseases across 8 organ systems, enabling evaluation of cross-domain diagnostic confusion errors.

Disease	EN	HI	BN	MR	PA	TE	AS
Asthma	63.2	42.1	0.0	73.7	0.0	15.8	0.0
Conjunctivitis	90.5	90.5	90.5	90.5	100.0	0.0	0.0
Coronary HD	63.2	68.4	73.7	47.4	14.8	26.3	0.0
Dermatitis	100	95.0	90.0	70.0	0.0	0.0	0.0
Enteritis	91.7	91.7	62.5	79.2	0.0	83.3	10.0
Esophagitis	81.5	70.4	81.5	74.1	0.0	14.8	0.0
Ext. Otitis	88.2	88.2	88.2	88.2	0.0	0.0	0.0
Mastitis	66.7	80.0	53.3	80.0	0.0	0.0	20.0
Pneumonia	45.0	20.0	0.0	40.0	70.0	55.0	0.0
Rhinitis	80.0	86.7	26.7	40.0	0.0	53.3	53.3
Thyroiditis	100	63.2	63.2	78.9	63.2	42.1	0.0
Brain Injury	94.7	94.7	73.7	73.7	0.0	0.0	0.0

Table 8: Per-disease post-processed accuracy (%) for `IndicMedLM` across selected languages. EN=English, HI=Hindi, BN=Bengali, MR=Marathi, PA=Punjabi, TE=Telugu, AS=Assamese. HD = Heart Disease. Zero entries indicate complete generation failure for that disease–language combination.

True Label	Predicted	Domain Shift	Risk
Coronary HD	Esophagitis	Cardiac→GI	**Critical**
Brain Injury	Esophagitis	Neuro→GI	**Critical**
Brain Injury	Enteritis	Neuro→GI	**Critical**
Pneumonia	Asthma	Resp→Resp	Moderate
Thyroiditis	Coronary HD	Endo→Cardiac	Moderate
Mastitis	Enteritis	Breast→GI	Moderate
Rhinitis	Pneumonia	Resp→Resp	Low
Conjunctivitis	Mastitis	Eye→Breast	Low

Table 9: Frequent cross-domain misclassifications for `IndicMedLM` with clinical risk stratification. HD = Heart Disease; Resp = Respiratory; Neuro = Neurological; Endo = Endocrine; GI = Gastrointestinal. Critical errors involve organ systems where misdiagnosis can cause irreversible harm.

Metric	Expert 1	Expert 2	Expert 3	Average
Medical Safety (Pass Rate)	96%	94%	96%	**95.3%**
Symptom Extraction	4.2	4.1	4.3	**4.20**
Context Memory	4.4	4.3	4.5	**4.40**
Diagnostic Correctness	4.1	4.0	4.2	**4.10**
Conversational Flow	4.3	4.2	4.4	**4.30**
Efficiency	4.0	3.9	4.1	**4.00**

Table 10: Medical expert evaluation of `IndicMedLM` across 50 sampled dialogues. Scores are reported on a 1–5 Likert scale except Medical Safety (Pass/Fail).

Original Disease	Misclassified As	Frequency
Pneumonia	Asthma	3
Esophagitis	Enteritis	2
Esophagitis	Asthma	2
Asthma	Pneumonia	2
Coronary heart disease	Asthma	2
Pneumonia	Enteritis	2
External otitis	Conjunctivitis	2
Conjunctivitis	Mastitis	2
Mastitis	Traumatic brain injury	2
Esophagitis	Coronary heart disease	1

Table 11: Most frequent disease-level misclassifications made by the final `IndicMedLM` model.

Metric	Krippendorff's α
Symptom Extraction	0.82
Context Memory	0.84
Diagnostic Correctness	0.80
Conversational Flow	0.83
Efficiency	0.78
Average	**0.81**

Table 12: IAA scores across three medical experts.

Prompt Type	Prompt Content
Translation Prompt	You are acting as a specialized Medical Translation Bridge, a critical link between an English-speaking doctor and a patient who speaks Assamese, Bengali, Gujarati, Hindi, Marathi, Punjabi, Tamil, Telugu, and Urdu. Your primary responsibility is to maintain absolute clinical accuracy while ensuring the tone is appropriately synced for both parties. When the doctor speaks in English, you must translate their advice, diagnoses, and prescriptions into the patient's native language using clear, empathetic, and culturally respectful terminology that a non-medical person can easily understand. Conversely, when the patient provides a query or describes symptoms in their native language, you will convert that input into precise, formal medical English for the doctor, ensuring that nuances of pain, duration, and history are preserved without loss of detail. You are strictly prohibited from hallucinating or adding medical advice not present in the source text, your role is purely to facilitate a perfectly synced, bidirectional exchange. Ensure that if the patient expresses distress or urgency, the English translation reflects that clinical priority to the doctor. Your output must contain only the translated text to allow for seamless integration into the communication interface.

Table 13: Prompt used for bidirectional medical translation in the multilingual inference layer.

Prompt Type	Prompt Content
Synthetic Dialogue Generation Prompt	Analyze `train.json` medical dialogues (patient/doctor exchanges, symptoms like "Cough", diagnoses such as "Esophagitis"). Create Python synthetic generator using Groq API (Llama-3 family model). Match exact format: `{'Dialog N': [{'patient': '...', 'doctor': '...'}]}`. Randomize symptom openings, generate 4–8 turns with doctor questions and realistic patient responses. Preserve the overall structure used for model training and provide progress, ETA, and resume-friendly execution. Output synthetic data in the same format as `train.json`.

Table 14: Prompt used to generate synthetic multi-turn medical consultations from the `MDDial` training distribution.

Prompt Type	Prompt Content
Dialogue Formatting Prompt	Convert a medical dialogue sample into ShareGPT-style multi-turn conversation. Structure: (1) the system message sets the medical diagnosis context, (2) patient utterances become `human` turns, (3) doctor utterances become `gpt` turns, and (4) the final `gpt` turn contains the diagnosis answer. Preserve dialogue order and ensure that each consultation remains a valid multi-turn interaction for instruction tuning.

Table 15: Prompt used to convert raw medical dialogues into ShareGPT-style training instances.

Evaluation Criterion	Description
Medical Safety (Pass/Fail)	Whether the system provides any potentially dangerous, misleading, or unsafe medical advice during the conversation.
Symptom Extraction (1–5)	Measures how accurately the model identifies and tracks the patient's symptoms throughout the dialogue.
Context Memory (1–5)	Evaluates whether the model remembers previously mentioned information such as symptoms or earlier responses in the conversation.
Diagnostic Correctness (1–5)	Assesses whether the final diagnosis is medically reasonable given the symptoms described in the conversation.
Conversational Flow (1–5)	Evaluates whether the dialogue is natural, coherent, empathetic, and professionally phrased, similar to a real clinical interaction.
Efficiency (1–5)	Measures whether the system asks an appropriate number of questions, avoiding unnecessary or redundant queries while still gathering sufficient information.
Annotator Notes	Free-text comments provided by medical experts to highlight issues such as reasoning errors, repeated questions, unsafe advice, or unusual dialogue patterns.

Table 16: Evaluation criteria used in expert assessment of the conversational medical system. Experts rated multiple aspects of safety, reasoning, and dialogue quality using a Likert scale (1–5), while medical safety was evaluated using a binary pass/fail metric.