# Can LLM Judges Reliably Distinguish Good and Bad Clinical Social Skills? A Persona-Grounded Benchmark in the Indian Clinical Context

**Dhruv Awasthi** **Shreya Gupta** **Anirudh Sharma**
Tulika Saha Dinesh Babu Jayagopi

International Institute of Information Technology, Bangalore, India

{dhruv.awasthi, shreya.gupta, anirudh.sharma, tulika.saha, jdinesh}@iiitb.ac.in

## Abstract

Doctor-patient communication involves more than medical correctness. Behaviours such as active listening, empathy, reassurance, and clear explanation play an important role in building patient trust and supporting clinical decision-making. However, most existing medical large language model (LLM) benchmarks focus primarily on factual reasoning and diagnostic capability, with limited evaluation of communication quality and behavioural consistency in multi-turn interactions. We introduce a persona-grounded benchmark for doctor-patient conversations in the Indian clinical context. We further develop a rubric-based evaluation framework with explicit four-level scoring and principle-adherence filtering across five dimensions: conversation initiation, responsiveness, empathy and emotional alignment, communication quality, and persona adherence. Using a synthetically generated benchmark of 150 persona-grounded doctor-patient conversations, we evaluate nine open-source and proprietary foundation models on their ability to distinguish desirable and intentionally poor clinical communication behaviours. The results show that several models assigned high behavioural scores even in conversations containing intentionally poor doctor personas, indicating difficulty in reliably separating socially aligned communication from persona-specific behaviour patterns. Overall, the benchmark highlights the importance of structured behavioural evaluation for clinical dialogue systems.

## 1 Introduction

Doctor-patient interaction is one of the main interfaces through which healthcare is delivered, and the *social skills* displayed in that interaction-how the doctor greets the patient, elicits symptoms, acknowledges emotions, explains procedures, handles consent, and closes the encounter-shape clinical outcomes, matters as much as the underlying medical knowledge. This is especially consequential in overburdened healthcare systems such as India, where short consultations, uneven access to specialists, varied levels of health literacy, and strong family involvement amplify the impact of social-skill failures on patient trust, adherence, and care experience.

Foundation models are becoming increasingly capable of producing fluent medical dialogue, and recent work has shown strong progress on medical question answering and diagnostic conversation systems (Singhal et al., 2023; Tu et al., 2025). A model that performs well on biomedical question answering, however, is not necessarily able to demonstrate clinically appropriate social skills when placed in the role of a doctor. Existing medical benchmarks largely measure factual recall and diagnostic reasoning. They do not directly evaluate whether a model can initiate a consultation, follow patient leads, acknowledge emotions, explain procedures in accessible language, handle consent in sensitive contexts, or maintain a specified clinical persona across a full interaction.

This gap matters for social-good applications of foundation models. Synthetic doctor-patient conversations underlie medical education, communication-skills training, dataset augmentation, red-teaming, and the development of clinical AI systems. These use cases require an evaluation framework that measures social skills directly-not as a byproduct of fluency or politeness. A model that is uniformly warm and helpful may appear safe on the surface, but if it cannot be evaluated on whether it greets the patient, reflects their concern back, or explains confidentiality in a sensitive setting, an important class of clinical failure modes remains invisible.

We address this gap with a benchmark for evaluating the social skills exhibited by foundation models in doctor-patient conversations, situated in the Indian clinical context. The benchmark is built

around a rubric-driven LLM-as-judge framework that scores doctor turns on five dimensions of clinical social skills: conversation initiation, responsiveness, empathy, communication accessibility, and adherence to a specified persona and context. Each dimension is operationalized through observable behavioral signals drawn from the clinical-communication literature, and several metrics incorporate a principle-adherence check that first determines whether a behavior is contextually applicable, so that models are not penalized for omitting empathy when no emotional cue is present or for not explaining confidentiality when the context is not sensitive.

To make this evaluation possible, we construct a test corpus of 150 synthetic doctor-patient conversations grounded in a few anonymized seed scripts from an Indian clinical setting, reviewed and validated by practicing Indian clinicians. The corpus pairs scenarios with desirable and undesirable doctor personas across twenty-one clinical conditions and diverse patient profiles. The corpus itself is not the primary contribution; it is the test bed that makes the social-skill evaluation possible. Earlier experiments with a GEval-style continuous LLM evaluator (Liu et al., 2023) produced high, poorly separated scores for conversations that differed meaningfully in social-skill quality, motivating the move to a categorical rubric with explicit decision boundaries.

The benchmark is organized around three research questions. **RQ1:** How reliably can LLM judges distinguish desirable from intentionally undesirable doctor social skills in clinical conversations? **RQ2:** Which dimensions of doctor social skills remain weak despite strong general fluency? **RQ3:** Can compact local models serve as reliable judges of clinical social skills, enabling cost-sensitive and privacy-sensitive evaluation in low- and middle-income clinical settings? These questions connect directly to the FMSG setting: the work concerns reliable and efficient foundation-model use for healthcare, with attention to cultural context, practical deployment constraints, and the risks of evaluating only surface-level language quality.

The main contributions are:

- **An evaluation framework for doctor social skills.** A rubric-driven LLM-as-judge framework that scores doctor-patient conversations along five behavioral dimensions of clinical social skills, using categorical 0-3 anchors and principle-adherence checks that prevent penalization when a behavior is not contextually applicable.
- **A benchmark of nine foundation models on doctor social skills.** A behavioral profile of where local and frontier systems succeed and where they fail, including a recurring conflict between helpfulness-tuned alignment and the faithful representation of undesirable clinical personas.
- **A localized Indian-context evaluation corpus.** 150 doctor-patient conversation scripts grounded in a clinician-reviewed seed, with paired desirable and undesirable doctor personas across twenty-one clinical conditions and diverse patient profiles, released to support future work on Indian clinical communication.
- **A methodological analysis of LLM-judge limitations in this setting.** Documentation of why GEval-style continuous LLM scoring proved insufficiently discriminative for clinical social-skill evaluation and how an anchored categorical rubric addresses the observed failure modes.

## 2 Related Work

### 2.1 Medical foundation models

Large language models have been evaluated extensively on medical knowledge tasks. Med-PaLM and MultiMedQA demonstrated that instruction-tuned LLMs can perform strongly on medical question-answering benchmarks and long-form consumer medical questions (Singhal et al., 2023). More recent conversational diagnostic systems such as AMIE focus directly on clinical dialogue, combining simulated interaction, automated feedback, and human evaluation across diagnostic and communication axes (Tu et al., 2025). Complementary efforts such as Healthcare Copilot further explore the use of general-purpose LLMs for end-to-end medical consultation workflows, highlighting their ability to structure multi-turn interactions while maintaining clinical relevance (Zhang et al., 2024).

Recent work has also emphasized the need for systematic evaluation of medical LLMs. A comprehensive taxonomy of evaluation methods cat-

egorizes benchmarks across knowledge, reasoning, safety, and interaction dimensions, highlighting gaps in dialogue-level assessment (Silva et al., 2025). Similarly, task-oriented evaluation frameworks such as MMDEval introduce interactive dialogue settings to assess model performance in realistic clinical tasks (Wang et al., 2025). Despite these advances, most systems remain focused on diagnostic correctness or task completion. The present work asks a complementary question: can foundation models exhibit clinically appropriate social skills, including in scenarios where the prescribed persona is intentionally undesirable, and can those social skills be evaluated at scale?

This distinction matters because medical competence and social competence are not identical. A model may recommend a plausible test while failing to ask an open-ended question, acknowledge fear, explain confidentiality, or check whether the patient understood the proposed procedure. Conversely, a model may sound warm and fluent while failing to maintain a deliberately undesirable persona needed for simulation. The proposed benchmark is therefore positioned closer to clinical-communication-skills evaluation than to factual medical QA.

### 2.2 Clinical communication assessment

Medical education has long treated communication as a trainable and assessable clinical skill. The Kalamazoo Consensus Statement identifies core physician communication tasks, including building the relationship, opening the discussion, gathering information, understanding the patient's perspective, sharing information, reaching agreement, and providing closure (Makoul, 2001). Protocols such as SPIKES similarly operationalize difficult communication tasks such as breaking bad news (Baile et al., 2000). These frameworks motivate the social-skill dimensions used in this work: initiation, responsiveness, empathy, accessible communication, and consent-related behaviours.

Recent benchmarks such as MedDialogRubrics and MEDPI extend this line of work by proposing structured evaluation frameworks for multi-turn medical consultations and patient-facing interactions, emphasizing rubric-based scoring and scenario diversity (Chen et al., 2026; Kumar et al., 2026). Unlike traditional checklists used in human training, these principles are adapted into an automated rubric for evaluating generated conversations.

The clinical-communication literature also motivates the emphasis on observable behaviours. Rather than asking whether a conversation is generally "high quality," the rubric asks whether the doctor greets the patient, invites the chief concern, uses open-ended questions, follows patient cues, reflects understanding, explains risks and benefits, and provides reassurance when appropriate. This makes the rubric closer to a clinical communication-skills checklist while remaining scalable through LLM-assisted scoring.

### 2.3 Persona-conditioned generation

Persona conditioning is commonly used to make dialogue agents consistent, controllable, and realistic. In healthcare simulation, persona fidelity is not merely stylistic. It determines whether a generated interaction can support training, evaluation, or red-teaming. Desirable doctor personas should demonstrate empathy, clarity, and patient-centredness, while undesirable doctor personas may display dismissiveness, bluntness, poor listening, or rushed consent. A model that collapses all personas into generic helpfulness can obscure clinically relevant failure modes.

Recent work such as Roleplay-doh enables domain experts to construct simulated personas grounded in explicit principles, improving controllability and realism in generated interactions (Louie et al., 2025). Similarly, selective prompting approaches for personalized conversations demonstrate how LLMs can be guided to maintain consistent behavioural traits across turns (Huang et al., 2025). These approaches reinforce the importance of controllable persona adherence, which is central to the social-skill evaluation proposed in this work.

### 2.4 LLM-as-judge evaluation

LLM-as-judge methods offer scalable reference-free evaluation and have been applied to summarization, dialogue, alignment, and open-ended generation (Liu et al., 2023; Li et al., 2024). However, prior work also notes reliability concerns, including bias, calibration instability, and weak reproducibility in some settings (Bavaresco et al., 2025). Preliminary GEval experiments in this project showed a domain-specific version of this problem: continuous numeric scoring saturated around moderate-to-high values and failed to separate conversations that were meaningfully different in clinical social-skill quality.

Recent simulation benchmarks such as AgentClinic further highlight the importance of structured evaluation in realistic clinical environments, combining multimodal interaction and task-driven assessment (Schmidgall et al., 2025). The framework developed here therefore uses explicit categorical rubrics and structured justifications rather than free numeric scores.

The approach does not claim that LLM judges replace expert clinical review. Instead, it treats the judge as a scalable first-pass evaluator whose decisions are constrained by explicit rubrics with anchored behavioral signals. This is especially useful during dataset and prompt iteration, where teams need to inspect many model outputs quickly. The final benchmark should still be strengthened through clinician validation, which remains an important next step.

## 3 Evaluation Corpus

### 3.1 Design Goals

The corpus was designed to support a rubric-based evaluation of doctor social skills in three respects. First, the conversations must contain enough interactional structure for the rubric's behavioral signals-greeting, paraphrasing, emotional acknowledgement, consent discussion-to be meaningfully present or absent. Second, the conversations must be localized to Indian clinical settings, where family involvement, respect for medical authority, home remedies, language mixing, and access constraints shape how patients behave and what doctors are expected to do. Third, the corpus must include both desirable and undesirable doctor personas in matched proportions, so that the same evaluation rubric can be applied across the social-skill spectrum without being driven by selection bias toward idealized examples.

These goals led to several construction constraints. Conversations had to be long enough for relationship-building and consent behaviours to appear, so very short single-turn exchanges were excluded. The patient was given enough agency to ask questions, hesitate, challenge, or defer to family, because such behaviours reveal whether the doctor follows patient leads or not. The doctor persona was required to influence the whole conversation rather than only the first utterance. Finally, the corpus had to include sensitive contexts-HIV testing, reproductive health, psychiatric treatment, genetic testing-where confidentiality and consent are not optional embellishments but core clinical behaviours.

| Axis | Examples |
|---|---|
| Patient persona | Fearful, skeptical, informed, reluctant, demanding, non-compliant |
| Doctor persona | Desirable: empathetic, supportive, investigative;<br>Undesirable: authoritative, blunt, pessimistic, busy |
| Condition | Cancer treatment, HIV testing, dialysis, anesthesia, blood transfusion |
| Diversity | Age, gender, economic status, health literacy, support system, urban/rural access |

Table 1: Corpus design axes used to construct persona-grounded doctor-patient conversations for evaluation.

### 3.2 Seed Script and Taxonomy

Corpus construction began from an anonymized seed consultation script obtained from a practicing Indian clinician. The seed focused on a minor procedure and included several behaviours relevant to social-skill evaluation: greeting, symptom elicitation, explanation of a procedure, risk-benefit discussion, patient hesitation, reassurance, consent, and closure. The seed was used as an interaction template, not as a memorized example. All identifying details were removed before further use, and the corpus consists entirely of synthetic conversations.

A taxonomy of patient and doctor profiles was constructed from project attribute sheets. Patient personas span emotional and interactional stances including fearful, frustrated, angry, embarrassed, overwhelmed, confused, skeptical, informed, reluctant, defensive, demanding, silent, non-compliant, and care-seeking. Doctor personas are divided into desirable and undesirable. Desirable doctor personas include empathetic, reassuring, supportive, investigative, patient-centred, optimistic, holistic, and detail-oriented styles. Undesirable doctor personas include authoritative, pessimistic, direct/blunt, and busy/rushed styles. Each conversation also samples a clinical condition (cancer treatment, tuberculosis treatment, kidney transplant, HIV/AIDS testing, blood transfusion, psychiatric treatment, reproductive health, cardiac procedures, dialysis, anesthesia, pain management, severe infections, among others) and a set of diversity attributes spanning age group, gender, economic status, health literacy, support system, and urban/rural access.

| Stage | Main refinement |
|---|---|
| v1-v3 | Seed-script prompting and manual prompt shaping for consent-oriented dialogue |
| v4-v7 | Larger script batches with randomized patient and doctor attributes |
| v8 | Hypothesis split and Western-to-Indian conversion experiment |
| v9-v11 | Explicit desirable/undesirable doctor branches and larger balanced script sets |
| v12-v13 | Small controlled regeneration passes for branch-specific debugging |
| v14 | Final 75/75 split with structured diversity sampling and persona-specific undesirable-doctor prompts |

Table 2: Summary of corpus-construction iterations.

The seed script and the generated corpus were reviewed by practicing Indian clinicians, who confirmed that the conversations were realistic to Indian clinical settings, captured culturally appropriate patient behaviours, and faithfully represented the prescribed doctor persona in both the desirable and undesirable branches. This clinician review is the basis on which the corpus is used as a reliable test bed for the social-skill evaluation that follows.

### 3.3 Corpus Construction Pipeline

The corpus was built across fifteen iterations of prompt design. Early iterations generated a small number of scripts from the seed consultation and focused on producing realistic consent-oriented doctor-patient dialogue. Subsequent iterations expanded the number of scripts, added random sampling from the attribute sheets, and introduced explicit patient persona, doctor persona, age, condition, and diversity fields. A middle phase experimented with cultural transfer: scripts were first generated in a Western clinical-communication style and then rewritten into an Indian context. This intermediate step helped identify which cultural elements should be represented directly in the prompt, including family consultation, home remedies, respectful address, fear of procedures, and doctor-as-authority dynamics. Later iterations removed the two-stage conversion and generated Indian-context conversations directly. A summary of the iterations appears in Table 2.

Several candidate models were considered for corpus generation, including GPT-4o-mini, GPT-4o, o1, and o1-mini. They were compared qualitatively during prompt development on their ability to produce natural multi-turn conversations, preserve the prescribed patient and doctor personas, maintain Indian cultural grounding, follow the desirable/undesirable doctor split, and scale to a 150-script corpus without excessive cost or latency. GPT-4o-mini was selected for the final corpus because it produced fluent conversations that preserved the requested metadata, consistently surfaced Indian-context cues such as family involvement and home-remedy references, and was more reliable for batch generation than the reasoning-oriented variants. This selection is reported for transparency about how the corpus was built; it is not a claim about which model is the best clinical-dialogue generator, and corpus generation is not the task that this paper evaluates.

The final phase split corpus construction into desirable and undesirable doctor branches. The desirable branch prompted for empathy, clarity, professionalism, accessible explanations, patient questions, and meaningful consent. The undesirable branch introduced persona-specific behaviour guides for authoritative, pessimistic, blunt, and busy doctors, instructing the generator to preserve these traits throughout the conversation. The final version also made diversity sampling more structured: instead of selecting a single broad diversity phrase, it sampled one value from each of six categories-age, gender, economic status, health literacy, support system, and location.

The final corpus contains 150 conversations: 75 with desirable-doctor personas and 75 with undesirable-doctor personas. Each conversation contains approximately 15 doctor turns and 15 patient turns, with metadata specifying the sampled patient persona, doctor persona, age, condition, and diversity profile. The corpus is synthetic and intended for research and evaluation only; it should not be treated as clinical advice.

### 3.4 Why Persona Contrast Matters for Evaluation

The 75/75 desirable/undesirable split is not an endorsement of poor clinical communication. It is an evaluation design choice. If all scripts in the corpus contained idealized doctors, the rubric and any LLM judge applied to it would tend to reward generic politeness, fluency, and medical relevance. Paired desirable and undesirable scenarios create a sharper test: the evaluation must distinguish a model that genuinely demonstrates patient-centred social skills from a model whose outputs are merely fluent, and it must also detect when a model has failed to maintain an undesirable persona that was

explicitly prescribed.

The undesirable branch also exposes a subtle alignment behaviour. Many instruction-tuned models have been optimized to avoid harmful or impolite responses. That objective is valuable for direct deployment, but it conflicts with educational and red-teaming use cases. A model that rewrites an authoritarian doctor into a supportive one may appear safer, yet it fails to demonstrate the persona-controlled behaviour that a training or red-team setting needs to study. For this reason, the evaluation treats persona fidelity as a social-skill dimension in its own right, separate from whether the underlying clinical behaviour is desirable.

## 4 Evaluation Framework

### 4.1 Evaluation Objectives

The evaluation framework was designed to assess behavioural aspects of clinical conversations rather than general response quality alone. The primary objective was to evaluate whether foundation models could reliably identify desirable and intentionally poor doctor behaviours across longer multi-turn interactions.

Unlike traditional dialogue evaluation settings that focus primarily on fluency, coherence, or factual correctness, our framework focuses on observable communication patterns within the consultation. These include behaviours related to empathy, responsiveness, conversational structure, patient engagement, and adherence to the intended doctor persona.

The framework was also designed to support interpretable scoring. Instead of relying on unrestricted quality judgments, evaluation was performed using explicit behavioural rubrics tied to clinically meaningful interaction patterns.

### 4.2 Behavioural Dimensions

The evaluation pipeline measures five high-level behavioural dimensions:

- **Conversation Initiation:** evaluates how the doctor begins the interaction, including greeting behaviour, introduction quality, patient acknowledgement, and consultation setup.
- **Responsiveness:** measures whether the doctor appropriately responds to patient concerns, emotional cues, follow-up questions, and contextual information shared during the interaction.
- **Empathy and Emotional Alignment:** evaluates emotional acknowledgement, reassurance, supportive communication, and the ability to respond appropriately to patient distress or anxiety.
- **Communication Quality:** focuses on clarity of explanation, patient-friendly language, engagement quality, conversational structure, and the ability to maintain meaningful interaction throughout the consultation.
- **Persona Adherence:** evaluates whether the doctor consistently maintains the intended behavioural persona across the conversation, including both desirable and intentionally poor communication styles.

| Dimension | Metrics |
|---|---|
| Initiation | Greeting, opening question, open-ended question |
| Responsiveness | Active listening, paraphrasing, following leads, topic redirection |
| Empathy | Personalization, emotion recognition, empathy, reassurance |
| Communication | Language complexity, fluency, confidentiality explanation |
| Persona | Doctor/patient persona and parameter adherence |

Table 3: Five-dimensional evaluation framework for clinical conversation quality.

These dimensions were further divided into 15 fine-grained behavioural metrics used during scoring, as shown in Table 3.

### 4.3 Rubric-Based Scoring

Initial experiments used GEval-style continuous scoring for behavioural assessment. However, this approach produced weak score separation across conversations with meaningfully different communication behaviours. Conversations that were fluent and coherent often received favorable scores even when important behavioural patterns were absent.

To improve interpretability and behavioural discrimination, the final framework uses explicit four-level behavioural rubrics for each metric. The scoring levels are shown in Table 4.

The rubric definitions were designed to prioritize observable behavioural cues instead of overall conversational fluency.

| Score | Interpretation |
|---|---|
| 0 | Behaviour absent or contradictory to expected interaction pattern |
| 1 | Limited or inconsistent behavioural alignment with noticeable communication deficiencies |
| 2 | Generally appropriate behavioural alignment with minor limitations |
| 3 | Strong and consistent behavioural alignment with intended interaction style |

Table 4: Four-level rubric used for behavioural evaluation.

| Model | Category |
|---|---|
| Llama 3.1 8B | Local 8B instruct model |
| Qwen3 8B | Local 8B Qwen-family model |
| Mistral 7B | Local 7B instruct model |
| Nemotron Nano 8B | Compact instruction-following model |
| Gemini 2.5 Flash | Frontier API model |
| Gemini 3 Flash | Frontier preview API model |
| Gemini 3.1 Pro | Frontier large-scale API model |
| DeepSeek Chat | Frontier conversational model |
| DeepSeek Reasoner | Reasoning-oriented frontier model |

Table 5: Foundation models evaluated using the proposed behavioural evaluation framework.

### 4.4 Principle-Adherence Filtering

Not all behavioural metrics are applicable to every clinical conversation. Certain interactions may not contain emotional distress, confidentiality concerns, or sensitive decision-making scenarios requiring evaluation of specific communication behaviours.

To address this, the evaluation pipeline includes a principle-adherence filtering mechanism. During scoring, non-applicable metrics are dynamically excluded from aggregation instead of being assigned low scores. In implementation, such metrics are represented using a sentinel value of `-1` before final score computation.

This prevents conversations from being penalized for behavioural dimensions that are irrelevant to the interaction context and improves fairness across diverse clinical scenarios.

### 4.5 Evaluation Pipeline

The proposed framework was used to evaluate nine open-source and proprietary foundation models as behavioural evaluators on the 150-conversation benchmark. The evaluated systems included both compact local models and larger frontier API models spanning different model families, scales, and alignment strategies.

Each evaluator model independently assessed all benchmark conversations using identical behavioural metrics, rubric definitions, and scoring instructions. The evaluation prompts explicitly defined the expected interaction behaviour for each metric together with the corresponding four-level rubric criteria in order to reduce ambiguity and improve scoring consistency across models.

During evaluation, models first identified the relevant portions of the interaction for a given metric and then assigned a categorical score based on observable behavioural evidence. Metrics that were not applicable to a particular conversation were excluded dynamically through the principle-adherence filtering mechanism described earlier.

The evaluated models are listed in Table 5.

Scores were aggregated at the metric, dimension, and model levels. In addition to overall aggregated scores, good-doctor and bad-doctor conversations were analyzed separately to evaluate persona preservation capability across different interaction settings.

The framework was designed to measure behavioural separation rather than binary good-versus-bad classification alone. Instead of assigning a single pass/fail label, the rubric provides a fine-grained behavioural profile across dimensions such as responsiveness, empathy, communication quality, and persona adherence. This enables more detailed analysis of evaluator strengths, behavioural biases, and failure modes across clinically relevant communication scenarios.

## 5 Results and Analysis

### 5.1 Overall Behavioral Separation

The results show a clear variation in how effectively different foundation models distinguish between desirable and intentionally poor doctor behaviors. Since all evaluated models act as behavioral evaluators rather than conversation generators, the primary objective was to measure how reliably each model could identify clinically meaningful interaction patterns and preserve behavioral separation across good-doctor and bad-doctor scenarios.

Overall, frontier proprietary models demonstrated strong performance in fluency-oriented and language-oriented metrics. However, several of these models also showed weaker separation between good and bad doctor personas, particularly when poor-doctor conversations remained conversationally coherent and medically informative.

In contrast, local models demonstrated comparatively stronger persona discrimination and behavioral consistency despite lower overall conversational sophistication.

### 5.2 Conversation Initiation and Responsiveness

Table 6 summarizes the results for conversation initiation metrics. Mistral 7B achieved the strongest overall performance across greeting quality, opening-question quality, and open-ended engagement, followed closely by Llama 3.1 8B and Qwen3 8B. Most other models showed moderate performance differences across these metrics.

Responsiveness-related metrics are shown in Table 7.

Qwen 3 8B, and Mistral 7B demonstrated strong performance in active listening and patient-following behaviour, followed by Llama 3.1 8B. However, paraphrasing quality remained inconsistent across several models, suggesting that reflective conversational grounding is harder to evaluate reliably than general conversational continuity. For Qwen3 8B, paraphrasing scores are omitted because the evaluator consistently marked the metric as "Not Applicable" across benchmark conversations.

### 5.3 Empathy and Communication Quality

Empathy-related results are shown in Table 8. Qwen 3 8B, Mistral 7B, and Llama 3.1 8B consistently achieved the strongest scores in emotional acknowledgement, reassurance, and supportive communication. Nemotron Nano 8B also demonstrated a better performance across empathy-oriented metrics against the proprietary models. Examining scores assigned to intentionally poor doctor conversations on Personalization specifically, Nemotron Nano 8B (2.25), Gemini 2.5 Flash (2.99), Gemini 3 Flash (2.99), Gemini 3.1 Pro (2.63), and DeepSeek Reasoner (2.12) assigned scores in the 2 to 3 range, while Llama 3.1 8B (0.88), Qwen3 8B (0.12), Mistral 7B (0.08), and DeepSeek Chat (1.60) assigned scores below 1.6 on the same conversations. The first group of judges fails to penalize undesirable doctor behavior on socially-loaded metrics such as personalization, empathy, and reassurance, even when the prescribed persona is intentionally poor. This suggests that highly aligned conversational models may sometimes prioritize socially supportive interpretation over strict persona-sensitive evaluation.

Communication-quality results are summarized in Table 9. Most frontier models achieved near-maximum fluency scores regardless of persona setting, indicating strong conversational smoothness across interactions. However confidentiality explanation showed substantially different behavior.

Overall, Qwen 3 8B consistently achieved strong scores in this set of metrics, followed by Nemotron Nano 8B and the proprietary models, receiving good scores. This indicates that conversational fluency alone does not necessarily correspond to a stronger evaluation of clinically relevant communication behaviors.

### 5.4 Persona Adherence

Persona adherence emerged as one of the most challenging dimensions overall. Table 10 shows that strong performance in empathy and fluency did not consistently translate into reliable evaluation of intentionally poor doctor personas.

Gemini 2.5 Flash and DeepSeek Chat achieved the strongest overall persona-adherence performance, while Nemotron Nano 8B demonstrated comparatively balanced behaviour across both good-doctor and bad-doctor settings.

In contrast, models such as Qwen3 8B and Mistral 7B showed substantial performance degradation in bad-persona evaluation, indicating weaker behavioural separation despite acceptable performance in supportive conversational settings.

### 5.5 Summary

The results indicate substantial variation in how different models evaluated clinically meaningful conversational behavior. The results indicate that strong conversational fluency and supportive language generation do not necessarily correspond to strong behavioral discrimination between appropriate and intentionally poor doctor interactions.

Open-weight models like Llama 3.1 8B, Qwen3 8B, and Mistral 7B demonstrated strong performance across initiation, responsiveness, and empathy-oriented metrics. Qwen3 8B achieved especially strong results in open-ended questioning, emotional recognition, empathy, reassurance, and active listening behaviours.

In contrast, proprietary models consistently achieved near-maximum fluency scores and highly polished conversational outputs. However, several of these models demonstrated weaker behavioral separation between good-doctor and bad-doctor

| Model | Greeting Quality | Opening Question | Open-Ended Questions |
|---|---|---|---|
| Llama 3.1 8B | 2.55 | 2.07 | 2.13 |
| Qwen3 8B | 2.34 | 1.89 | 2.51 |
| Mistral 7B | 2.69 | 2.05 | 2.27 |
| Nemotron Nano 8B | 1.67 | 1.37 | 1.95 |
| Gemini 2.5 Flash | 1.51 | 1.56 | 1.39 |
| Gemini 3 Flash | 1.43 | 1.54 | 1.48 |
| Gemini 3.1 Pro | 1.50 | 1.54 | 1.50 |
| DeepSeek Chat | 1.55 | 1.53 | 1.49 |
| DeepSeek Reasoner | 1.48 | 1.56 | 1.47 |

Table 6: Model performance evaluated using metrics covering conversation initiation.

| Model | Active Listening | Paraphrasing | Following Leads | Topic Redirection |
|---|---|---|---|---|
| Llama 3.1 8B | 2.29 | 1.82 | 2.13 | 2.01 |
| Qwen3 8B | 2.82 | NA | 2.82 | 2.00 |
| Mistral 7B | 2.65 | 2.00 | 2.14 | 2.46 |
| Nemotron Nano 8B | 1.73 | 1.41 | 1.62 | 1.35 |
| Gemini 2.5 Flash | 1.51 | 1.40 | 1.57 | 1.63 |
| Gemini 3 Flash | 1.50 | 1.51 | 1.51 | 1.00 |
| Gemini 3.1 Pro | 1.47 | 1.51 | 1.48 | 0.00 |
| DeepSeek Chat | 1.59 | 1.46 | 1.55 | 1.00 |
| DeepSeek Reasoner | 1.50 | 1.41 | 1.46 | 0.00 |

Table 7: Performance of various models across metrics covering responsiveness of a conversation.

personas, particularly when poor-doctor conversations remained conversationally coherent or medically informative. As summarized in Table 11, local models, particularly Qwen3 8B, outperformed several proprietary models, likely due to their stronger adherence to explicit behavioural cues and reduced tendency toward alignment-driven reasoning that smooths undesirable personas.

These findings suggest that behavioral evaluation depends on more than general conversational quality alone. Models optimized for fluency and socially supportive interaction may struggle to consistently identify subtle but clinically important negative conversational behaviors. Empathy-focused evaluation settings may benefit from frontier conversational models, whereas simulation-oriented applications requiring reliable negative-persona assessment may benefit more from models demonstrating stronger behavioral separation and persona adherence consistency.

## 6 Discussion

**Implications for Indian clinical settings.** The benchmark emphasizes culturally situated social skills rather than generic medical helpfulness. Indian clinical interactions can involve family consultation, deference to doctors, fear of procedures, reliance on home remedies, and uneven access to care. Models that ignore these factors may still sound fluent but fail to reproduce the social-interaction context that trainees or evaluators need. Evaluating doctor social skills in this context is therefore not a stylistic preference; it is a precondition for trustworthy use of foundation models in Indian healthcare applications.

**Helpfulness alignment can reduce social-skill controllability.** The most striking finding is that models often resist undesirable-doctor behaviour by making the doctor warmer, more polite, or more supportive than prescribed. This may be desirable for direct patient-facing deployment, but it is a limitation for simulation, red-teaming, and training-data generation. Social-good uses of foundation models require the ability to exhibit harmful or suboptimal social-skill patterns under controlled conditions so that they can be detected, studied, and mitigated.

**Rubric design materially changes what evaluation can reveal.** The move from GEval-style continuous scoring to anchored categorical rubrics changed what the evaluation could reveal about doctor social skills. Continuous LLM-judge scores rewarded fluency and produced weak discrimination across social-skill quality. The anchored rubric exposed concrete failure modes: missing greetings, absent paraphrasing, weak confidentiality explanations, and persona collapse in undesirable sce-

| Model | Personalization | Emotion Recognition | Empathy | Reassurance |
|---|---|---|---|---|
| Llama 3.1 8B | 1.99 | 2.28 | 2.29 | 2.65 |
| Qwen3 8B | 2.75 | 2.61 | 2.94 | 2.96 |
| Mistral 7B | 2.49 | 1.95 | 2.46 | 2.70 |
| Nemotron Nano 8B | 1.57 | 2.17 | 1.82 | 1.59 |
| Gemini 2.5 Flash | 1.51 | 1.63 | 1.51 | 1.51 |
| Gemini 3 Flash | 1.51 | 1.54 | 1.49 | 1.50 |
| Gemini 3.1 Pro | 1.47 | 1.59 | 1.53 | 1.50 |
| DeepSeek Chat | 1.57 | 1.51 | 1.52 | 1.57 |
| DeepSeek Reasoner | 1.54 | 1.57 | 1.53 | 1.50 |

Table 8: Comparison of model performance on metrics across empathy and communication quality.

| Model | Language Complexity | Fluency | Confidentiality Explanation |
|---|---|---|---|
| Llama 3.1 8B | 1.89 | 1.88 | 1.98 |
| Qwen3 8B | 2.44 | 2.80 | 1.53 |
| Mistral 7B | 2.45 | 1.30 | 1.58 |
| Nemotron Nano 8B | 1.73 | 2.98 | 1.58 |
| Gemini 2.5 Flash | 1.53 | 2.99 | 1.53 |
| Gemini 3 Flash | 1.51 | 3.00 | 1.51 |
| Gemini 3.1 Pro | 1.43 | 3.00 | 1.52 |
| DeepSeek Chat | 1.54 | 2.99 | 1.54 |
| DeepSeek Reasoner | 1.61 | 2.99 | 1.52 |

Table 9: Overall communication-quality scores across evaluated models.

| Model | Persona Adherence |
|---|---|
| Llama 3.1 8B | 1.80 |
| Qwen3 8B | 1.81 |
| Mistral 7B | 1.18 |
| Nemotron Nano 8B | 1.95 |
| Gemini 2.5 Flash | 2.28 |
| Gemini 3 Flash | 1.39 |
| Gemini 3.1 Pro | 1.09 |
| DeepSeek Chat | 2.15 |
| DeepSeek Reasoner | 1.93 |

Table 10: Table of model score on the persona-adherence metric.

| Model | Good | Bad | Gap (G−B) |
|---|---|---|---|
| Qwen3 8B | 2.35 | 0.52 | **1.83** |
| Mistral 7B | 1.87 | 0.28 | 1.59 |
| Llama 3.1 8B | 2.08 | 0.85 | 1.23 |
| Nemotron Nano 8B | 2.65 | 2.36 | 0.29 |
| Gemini 2.5 Flash | 2.19 | 2.12 | 0.07 |
| DeepSeek Chat | 1.72 | 1.66 | 0.06 |
| DeepSeek Reasoner | 1.83 | 1.83 | 0.00 |
| Gemini 3 Flash | 2.05 | 2.01 | 0.04 |
| Gemini 3.1 Pro | 1.92 | 1.95 | −0.03 |

Table 11: Judge discrimination on the persona-grounded benchmark. **Good** and **Bad** are mean scores assigned by each judge to desirable-doctor and undesirable-doctor conversations, respectively. **Gap (G−B)** is the discrimination gap: a higher gap indicates a judge that more reliably separates desirable from intentionally undesirable clinical social skills. A near-zero or negative gap indicates a judge that fails to discriminate.

narios. This suggests that domain-specific LLM-as-judge evaluation should be grounded in explicit behavioural signals rather than generic quality judgments.

**Ethical considerations.** The corpus is synthetic and should not be used as clinical advice. Undesirable-doctor scripts are included only for research, simulation, and safety evaluation. Any release will document intended uses and prevent presentation of poor doctor behaviour as acceptable practice. The seed script was anonymized, and no protected health information appears in the released data. For double-blind submission, repository links and institutional identifiers remain anonymized.

**Limitations.** The corpus contains 150 scripts and is not a comprehensive representation of Indian healthcare. The conversations are derived from one seed interaction and a finite taxonomy of personas and conditions. The evaluation relies on an LLM judge, which can carry its own biases despite the anchored rubric. Formal clinician-rater validation of the judge's social-skill scores is ongoing and should be added before using these metrics in deployment-facing settings. The benchmark is currently English/Hinglish-oriented and does not cover the full multilingual diversity of Indian healthcare. Per-conversation, per-judge raw score distributions are released alongside the corpus and

prompts at the project repository, enabling independent inspection of judge behavior on individual conversations and dimensions.

## 7 Conclusion and Future Work

This paper introduced a rubric-driven LLM-as-judge benchmark for evaluating the social skills exhibited by foundation models in doctor-patient conversations, grounded in the Indian clinical context. Across nine local and frontier foundation models, the evaluation reveals that systems are often fluent and empathetic but inconsistently reliable on structured clinical social skills such as initiation, paraphrasing, confidentiality explanation, and undesirable-persona discrimination. Qwen3 8B is among the most discriminating judges of clinical social skills despite its compact size, suggesting that smaller open-weight models can be viable for cost-sensitive and privacy-sensitive clinical evaluation. The rubric definitions, prompts, evaluation outputs, and corpus are released at `https://github.com/DhruvAwasthi/DocPersona-IN` to support reproducibility. Future work will scale the evaluation corpus, add formal clinician-rater validation of judge scores, expand to Indian languages, and evaluate multilingual and speech-based interactions.

## Acknowledgments

We thank the clinicians who reviewed the seed script and validated the corpus.

## A Evaluation Logs

To enable independent inspection of the judge behavior described in Section 4, the full evaluation logs containing one entry per (judge, metric, conversation) tuple are released at `https://github.com/DhruvAwasthi/DocPersona-IN`.

## References

Walter F. Baile, Robert Buckman, Renato Lenzi, Gary Glober, Estela A. Beale, and Andrzej P. Kudelka. 2000. SPIKES: A six-step protocol for delivering bad news: Application to the patient with cancer. *The Oncologist*, 5(4):302–311.

Andrea Bavaresco, Raffaella Bernardi, Federico Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Alexander Koller, Benjamin Newman, Sandro Pezzelle, Barbara Plank, David Schlangen, and 1 others. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. *Transactions of the Association for Computational Linguistics*. Also available as arXiv:2406.18403.

L. Chen and 1 others. 2026. Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint arXiv:2601.03023*.

Qiushi Huang and 1 others. 2025. Selective prompting tuning for personalized conversations with large language models. *arXiv preprint*.

A. Kumar and 1 others. 2026. Medpi: Evaluating ai systems in medical patient-facing interactions. *arXiv preprint arXiv:2601.04195*.

Dawei Li and 1 others. 2024. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Ryan Louie and 1 others. 2025. Roleplay-doh: Enabling domain experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint*.

Gregory Makoul. 2001. Essential elements of communication in medical encounters: The Kalamazoo consensus statement. *Academic Medicine*, 76(4):390–393.

Samuel Schmidgall and 1 others. 2025. Agentclinic: A multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint*.

J. Silva and 1 others. 2025. Evaluation of medical large language models: Taxonomy, review, and directions. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathanael Scharli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, and 7 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

Tao Tu, Arvind Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Vivek Natarajan, Juraj Gottweis, Yun Liu, and S. Sara Mahdavi. 2025. Towards conversational diagnostic artificial intelligence. *Nature*. Also available as arXiv:2401.05654.

X. Wang and 1 others. 2025. Interactive evaluation for medical llms via task-oriented dialogue system. In *Proceedings of COLING*.

Y. Zhang and 1 others. 2024. Healthcare copilot: Eliciting the power of general llms for medical consultation. *arXiv preprint arXiv:2402.13408*.