PRISM: Physician Rules Integrated with Small Language Models for Probable Diagnoses Associated with Abdominal Pain

Gautam Ahuja[1,2,3,4,#] Ayush Agarwal[5,#] Hara Prasad Mishra[2,4,6,#] Samagra Agrawal[5] Rik Ganguly[3,4] Zonunmawia[2,4,7] **Akshay Sharma**[2,4,8] **Vatsal Batra**[2,4,9] Bableen Kaur[2,4] Siddhant Poudyal[2,4] Himani Balutia[2,4] Sagarika[2,10] Sanjana Ahuja[2,4] Kedar Natarajan[2,11] Partha Pratim Das[1] Ramesh Jain[1,12] Partha Pratim Chakrabarti[13] Anurag Agrawal[2,10,14,†] Govind K Makharia[5,†] Rintu Kutum[1,2,3,4,14,†,]

Koita Centre for Digital Health at Ashoka | Mphasis AI & Applied Tech Lab at Ashoka

Article | Lab

HIGHLIGHTS

- **Real-World Deployment**: The PRISM is deployed as a functional Android mHealth app featuring Voice-to-Text and Text-to-Speech (Kokoro 82M) for accessible patient interaction.
- **Privacy-First Edge Computing**: The PRISM runs entirely on consumer-grade GPUs (< 14B parameters), ensuring that patients' data remains local and secure.
- **Resource Efficiency**: Qwen-2.5 (1.5B & 14B) maximizes performance, achieving high empathy and keyword extraction scores with minimal GPU memory footprint (< 4GB & < 14GB).
- **Physician-centric and Physician-in-the-Loop**: Unlike "black box" AI, every probable diagnosis is traceable via deterministic clinical rules, ensuring clinical safety and explainability.

PROBLEM & MOTIVATION

Abdominal pain is one of the most common yet a significant diagnostic challenge in high-volume, resource-constrained outpatient departments. At the All India Institute of Medical Sciences (AIIMS), New Delhi, India, the Gastroenterology OPD alone managed ≈ 0.11 million cases in 2023-2024. Abdominal pain spans numerous organ systems and often requires multiple diagnostic modalities to be considered before concluding the final diagnosis. End-to-end large language models (LLMs) offer conversational flexibility but are prone to hallucinations and lack the reliability and transparency required for clinical diagnosis. On the other hand, traditional rule-based systems are transparent & safe, but rigid and lack natural conversational abilities. Our approach (**PRISM**) is a Hybrid Conversational System (hCS) that combines the safety of physician-guided, rules-based logic for diagnostic reasoning with an ensemble of open-source small language models (SLMs) for patient-facing conversational and targeted non-diagnostic tasks.

CLINICAL PROTOCOL & WORKFLOW

Physician-Curated Logic: The core diagnostic logic is a physician-curated knowledge base, developed and validated by three gastroenterologists. **Workflow & Triage:** The system utilizes a structured questionnaire comprising 17 questions, yielding 29 potential diagnoses and 19 identified organs of origin. It includes discriminator questions to distinguish emergency cases from regular OPD cases, which branches further as shown in Figure 1C.

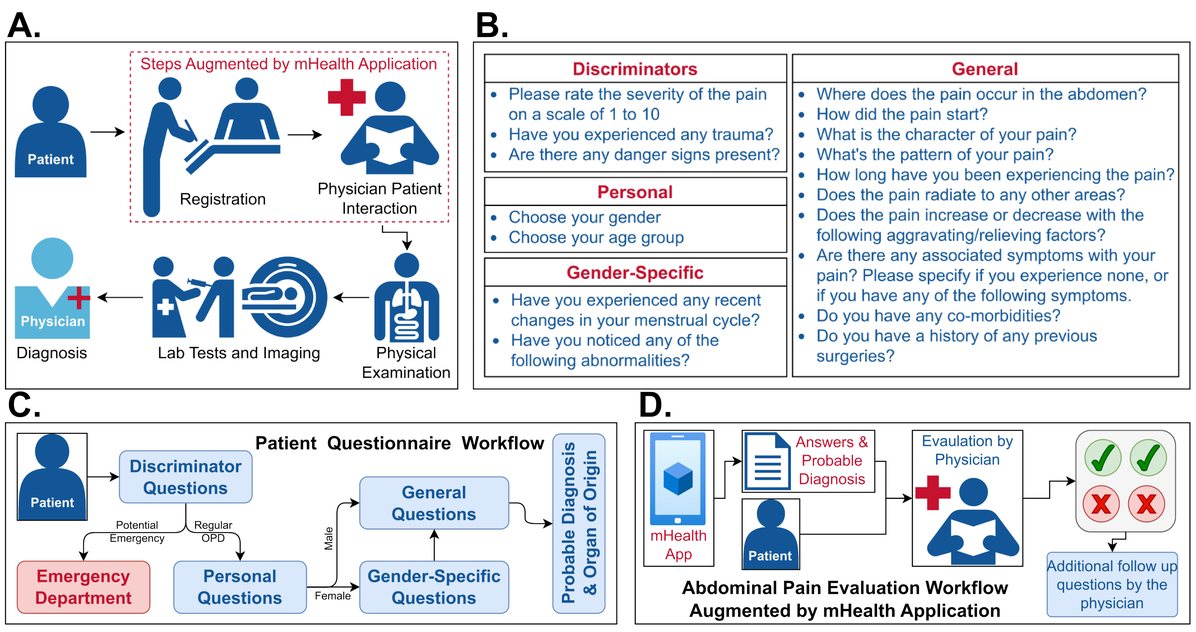

Figure 1. **Physician-Centric Protocol & Workflow.** **(A)** PRISM augments the physician-patient interaction phase to provide a list of probable diagnoses and reduce OPD burden without replacing the expert. **(B)** The 17 item questionnaire is strictly physician-curated, categorizing inquiries into discriminators, personal, gender-specific and general symptoms. **(C)** Built-in Triage Logic automatically filters and redirects emergency cases (e.g., trauma, danger signs) before the regular OPD interaction proceeds. **(D)** The clinical validation workflow, where physicians cross-verify the app's diagnosis against standard care to ensure safety.

SYSTEM ARCHITECTURE

PRISM employs a **modular microservice-based architecture** where an *Orchestrator* manages independent services: *Conversational*, *Empathizer*, *Keyword Extractor*, and *Mapper*. It is designed to run on-premise using open-source SLMs without external data calls, ensuring data security and privacy.

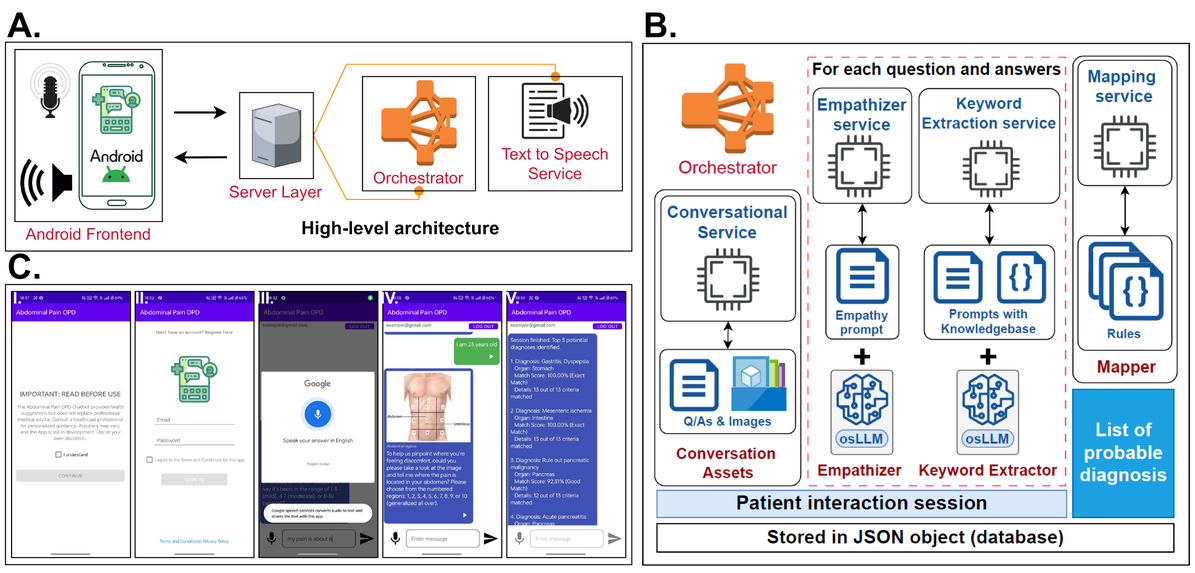

Figure 2. **Hybrid System Architecture & Implementation.** **(A)** High-level overview of the privacy-first, on-premise client-server interaction. **(B)** The Microservices Backend combines the best of both worlds: SLMs handling empathy and extraction, coordinated with a Deterministic Mapper for safe, rule-based diagnosis. The conversations are stored as JSON objects with random UUIDs. **(C)** The deployed Android mHealth App shows multimodal voice-to-text interaction, secure login, and transparent, ranked probable diagnoses for the physician.

MULTISTEP KNOWLEDGE INJECTION

To convert vague patient descriptions (e.g., "feels like a fire") into structured data that rigid rule engines can process, we expanded the initial 76 physician-curated keywords into $3,800$ **standardized terms** across 13 **vocabularies** (including SNOMED-CT and MTH) to capture diverse semantic variations using UMLS. A dynamic injection strategy that minimizes context noise and hallucinations:

- **Step 1 (Filter):** The SLM analyzes the narrative to identify relevant high-level symptom categories (e.g., "Burning Pain").
- **Step 2 (Extract):** The system injects *only* the specific synonyms for those categories into the final prompt, strictly grounding the model's output.

KEY RESULTS

We conducted a series of experiments for the services in PRISM, including two services that utilize SLMs (*Empathizer* and *Keyword Extractor*), as well as ranking methods for the mapper service. We benchmarked the performance of the keyword extractor using **SynD1** (Synthetic Dataset 1), consisting of 60 patient interaction-like question-answer pairs across varying difficulty levels generated using Gemini 2.5 Pro. We also utilized **SimD2** (Simulated Dataset 2), a dataset of 20 simulated patient scenarios explicitly curated by Gastroenterologists and Clinical Pharmacologists, which was used for the evaluation of the SLM in keyword extraction and mapping the top 5 probable diagnoses to rigorously test clinical validity. These 20 clinical scenarios contain a total of 322 question-answer pairs.

Empathy Benchmarking

- The rating was performed in three dimensions: empathy, clarity, and helpfulness by the nine human raters (non-physicians and pharmacologists).
- Raters scored the 20 randomized paraphrased versions of each of the 29 questions. The top 5 results are shown in Table 1.
- The Qwen-2.5 (1.5B) model achieved the highest human ratings for empathy (3.82 ± 0.09) and helpfulness (4.05 ± 0.09), proving massive models are not required for patient rapport.

Table 1. Benchmark of SLMs for Empathizer Service (Top 5 models shown).

Base model	# parameters (in billions)	Empathy	Clarity	Helpfulness	Overall
Qwen-2.5	1.5	3.82 ± 0.09	4.09 ± 0.08	4.05 ± 0.09	3.99 ± 0.08
Qwen-2.5	14	3.72 ± 0.09	4.12 ± 0.08	3.95 ± 0.08	3.93 ± 0.08
Llama-3.2	1	3.77 ± 0.08	4.02 ± 0.09	3.99 ± 0.09	3.92 ± 0.09
Llama-3.2	3	3.72 ± 0.09	4.10 ± 0.08	3.90 ± 0.09	3.90 ± 0.09
Gemma-2	9	3.67 ± 0.10	4.11 ± 0.08	3.93 ± 0.10	3.90 ± 0.09

Keyword Extraction Benchmark

- In the SynD1, Qwen-2.5 (14B) and Gemma-3 (12B) outperformed with an F1 score of 1.0 and 0.91 in the baseline and baseline with UMLS terms settings, respectively, and Phi-4 (14B) reasoning model and Gemma-3 (12B) scored an F1 score of 1.0 in baseline with targeted knowledge injection. The results are shown in Figure 3A.
- In the SimD2 dataset, Qwen-2.5 (14B) outperformed all other models in all three settings with an F1 score of 0.99 (baseline), 0.94 (baseline + UMLS terms), and 0.98 (baseline + targeted knowledge injection). The results are shown in Figure 3B.

Diagnostic Accuracy (Hybrid vs. End-to-End):

- PRISM (Hybrid): Achieved Top-5 Accuracy of 80% (MRR 0.596; V1 Mapper) – 100% (MRR 0.603; V2 Mapper) on SimD2 dataset.
- Pure LLMs: The best end-to-end model (Phi-4 Reasoning) only reached 65% (MRR 0.422) accuracy.
- The Top5 accuracy and MMR results are shown in Figure 3 C and D, respectively.

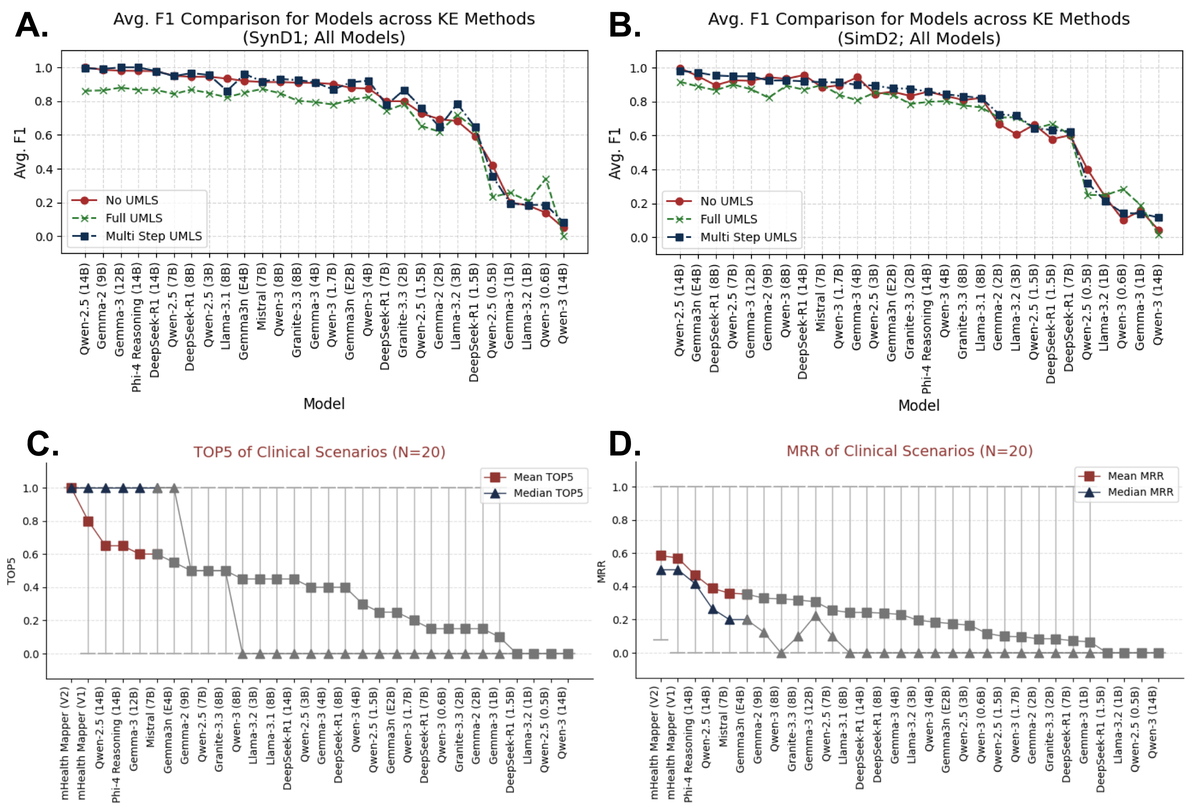

Figure 3. **Key Results - Hybrid vs. End-to-End.** **(A & B)** Keyword extraction performance (F1-score) on SynD1 and SimD2 datasets; Qwen-2.5 (14B) achieves near-perfect extraction. **(C & D)** Diagnostic accuracy comparing PRISM against pure end-to-end SLMs. PRISM (mHealth Mapper V2) achieves 100% Top-5 Accuracy and superior Mean Reciprocal Rank (MRR; 0.6), significantly outperforming the best standalone GenAI models (e.g., Phi-4, Qwen-2.5).

FUTURE DIRECTIONS

- **Adversarial Robustness:** Conducting perturbation testing and vulnerability assessments to ensure SLM stability under noisy or adversarial inputs.
- **Extended Testing:** Testing the performance on a larger physician-curated simulated dataset.
- **Multilingual Accessibility:** Integration of Indic models (e.g., AI4Bharat) to create a seamless end-to-end pipeline: *Indic Voice* → *English Processing* → *Indic Response*.
- **Clinical Validation:** Planning a **Randomized Clinical Trial (RCT)** at AIIMS, New Delhi, to rigorously evaluate diagnostic validity and system robustness in a live OPD setting.
- **Knowledge Standardization:** Enriching the semantic framework by integrating with the **Observational Medical Outcome Partnership (OMOP) Common Data Model (CDM)** from Observational Health Data Science and Informatics (OHDSI), and expanding into indigenous Indic health knowledge bases.

Affiliations: [1]Department of Computer Science, Ashoka University, India; [2]Koita Center for Digital Health at Ashoka (KCDH-A), Ashoka University, India; [3]Mphasis AI & Applied Tech Lab at Ashoka, Ashoka University, India; [4]Augmented Health Systems Laboratory, Ashoka University, India; [5]Department of Gastroenterology & Human Nutrition, AIIMS, New Delhi, India; [6]Department of Pharmacology, University College of Medical Sciences, New Delhi, India; [7]Department of Computer Science, International University of Applied Sciences, Bad Honnef, Germany; [8]Royal Bank of Canada, Toronto, Canada; [9]University of Massachusetts Amherst, Massachusetts, USA; [10]Department of Biology, Ashoka University, India; [11]University of Southampton, Southampton, UK; [12]UCI Institute for Future Health & University of California, Irvine, USA; [13]Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India; [14]Trivedi School of Biosciences, Ashoka University, India; [#]Equal contributor; [†]Co-correspondence: anurag.agrawal@ashoka.edu.in, govindmakharia@gmail.com; [*]Lead correspondence: rintu.kutum@ashoka.edu.in, rintu.kutum@augmented-health-systems.org