

Koita Centre for Digital Health Annual Day 2025

Conversational mHealth app with language model and a rule-based decision support system for probable diagnosis of abdominal pain in an OPD setting

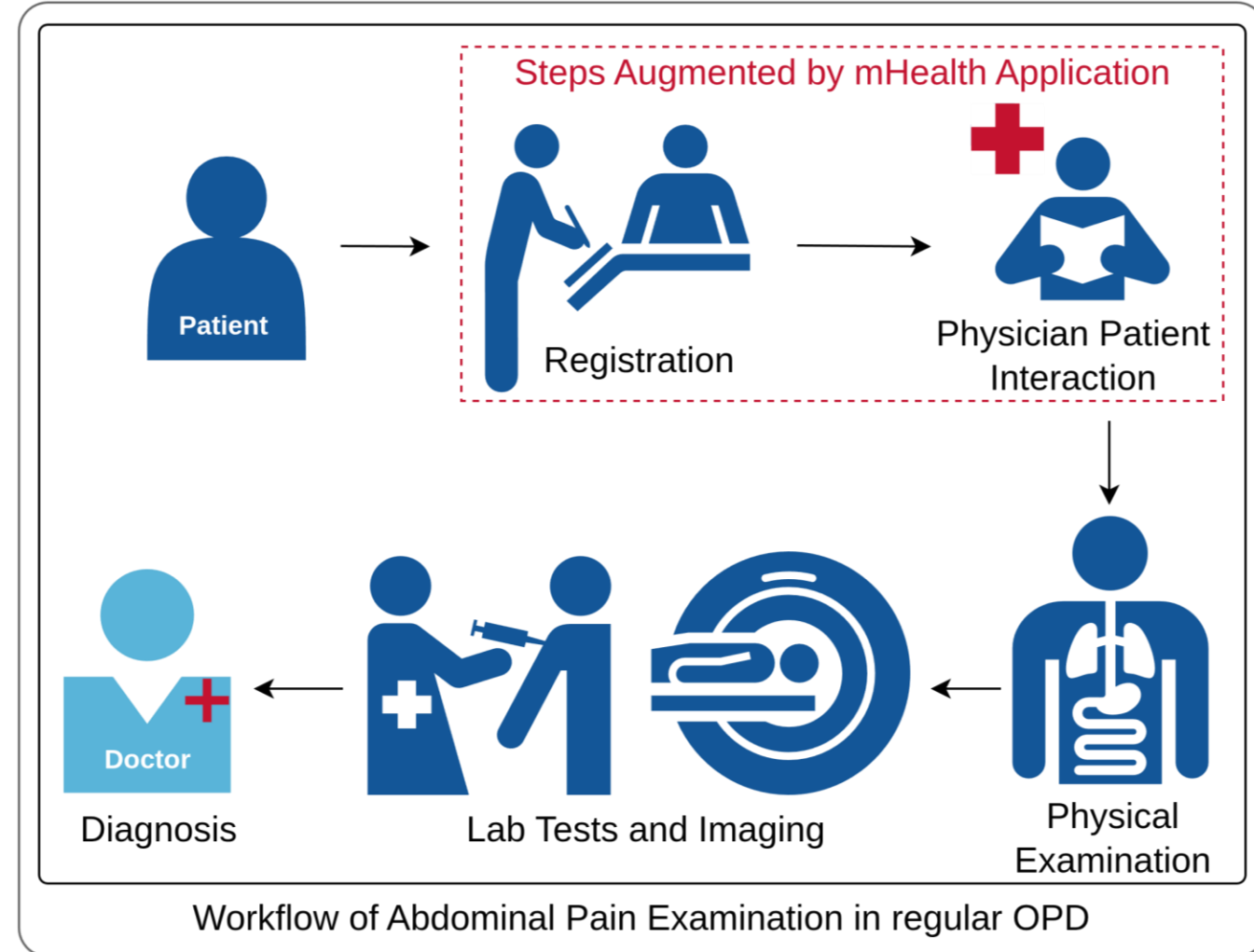
Gautam Ahuja

CS Major & Math Conc., ASP25

KCDH-A Research Intern (2024-25)

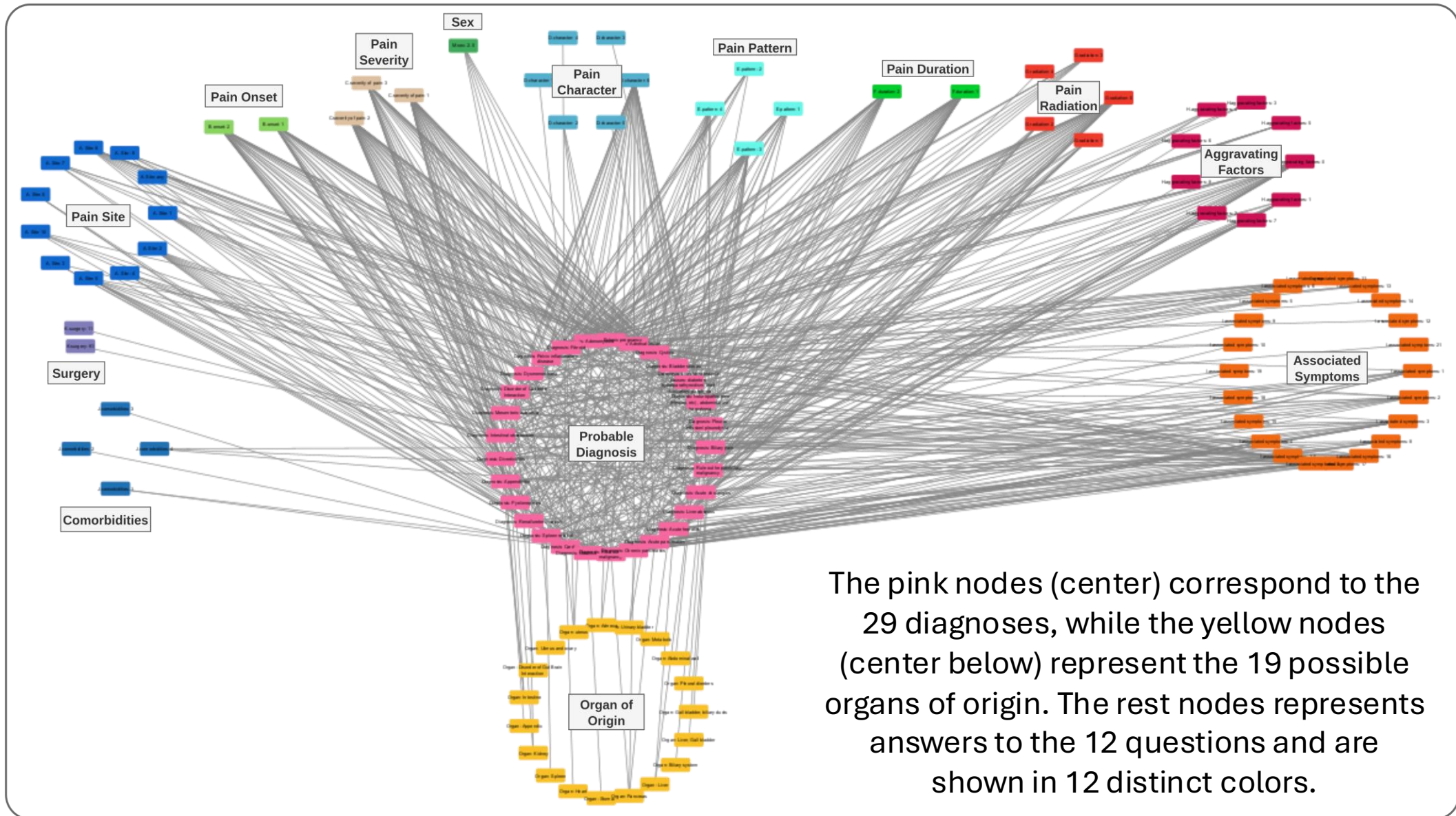
Background

- **Large volume of patients:** The Department of Gastroenterology and Human Nutrition at AIIMS New Delhi managed **1,04,544** out of **21,72,531** outpatient department cases at the main hospital in 2023-2024*.
- **Physician burden:** Physicians at gastroenterology will evaluate approximately 400 patients between 8:30 a.m. to 1:00p.m. everyday.
- Typical workflow of diagnosing the complaint of abdominal pain starts with registration and a physician-patient interaction. The **mHealth application will augment the initial step by identifying probable diagnoses and their organ of origin** for the physicians to consider.



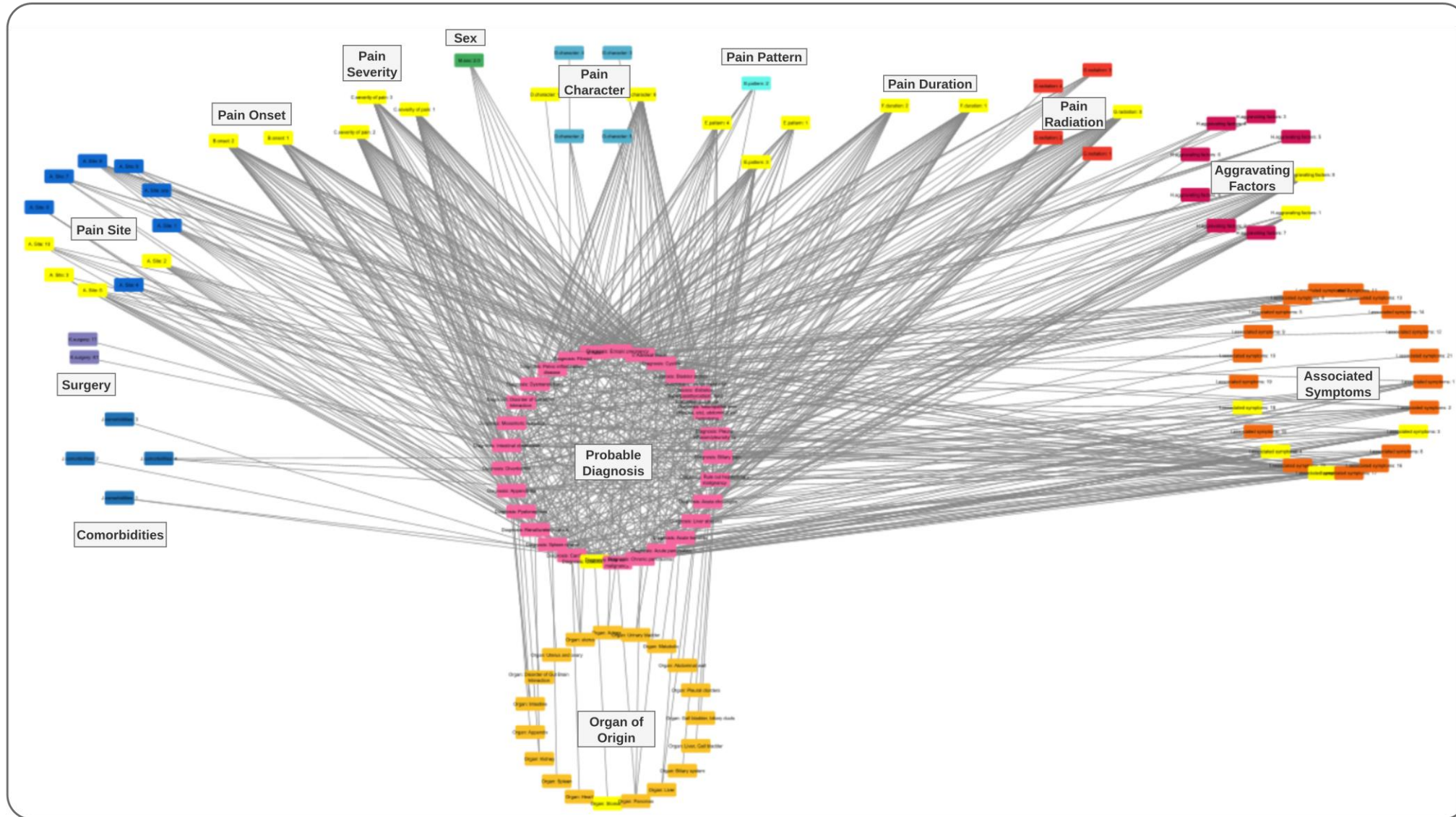
[*] 68th AIIMS Annual Report 2023-2024

Probable Diagnosis and Organ of Origin Mapping

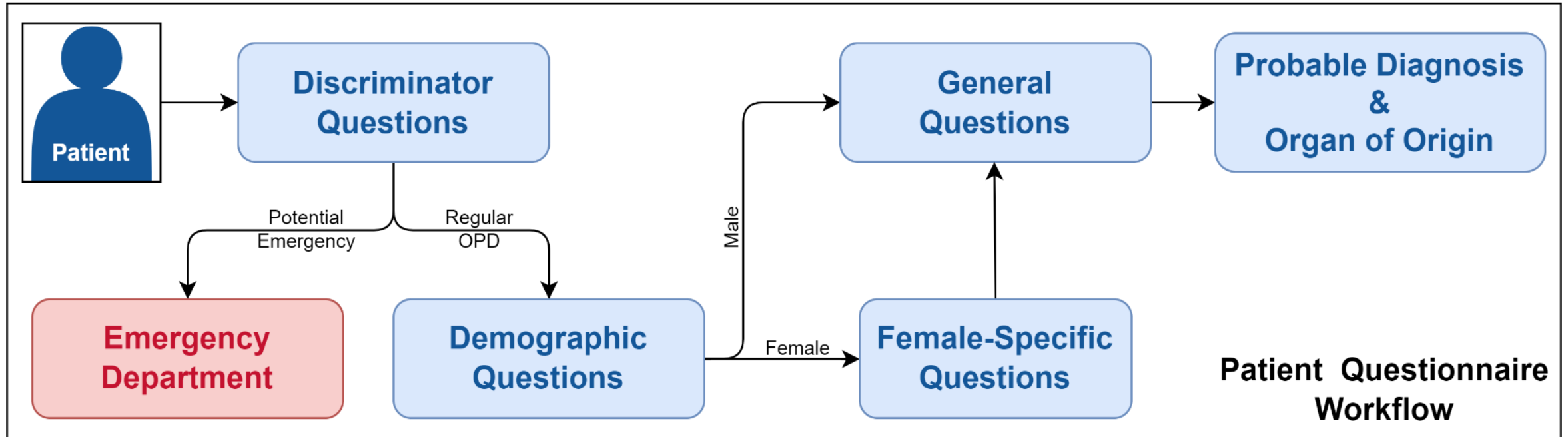


Probable Diagnosis and Organ of Origin Mapping

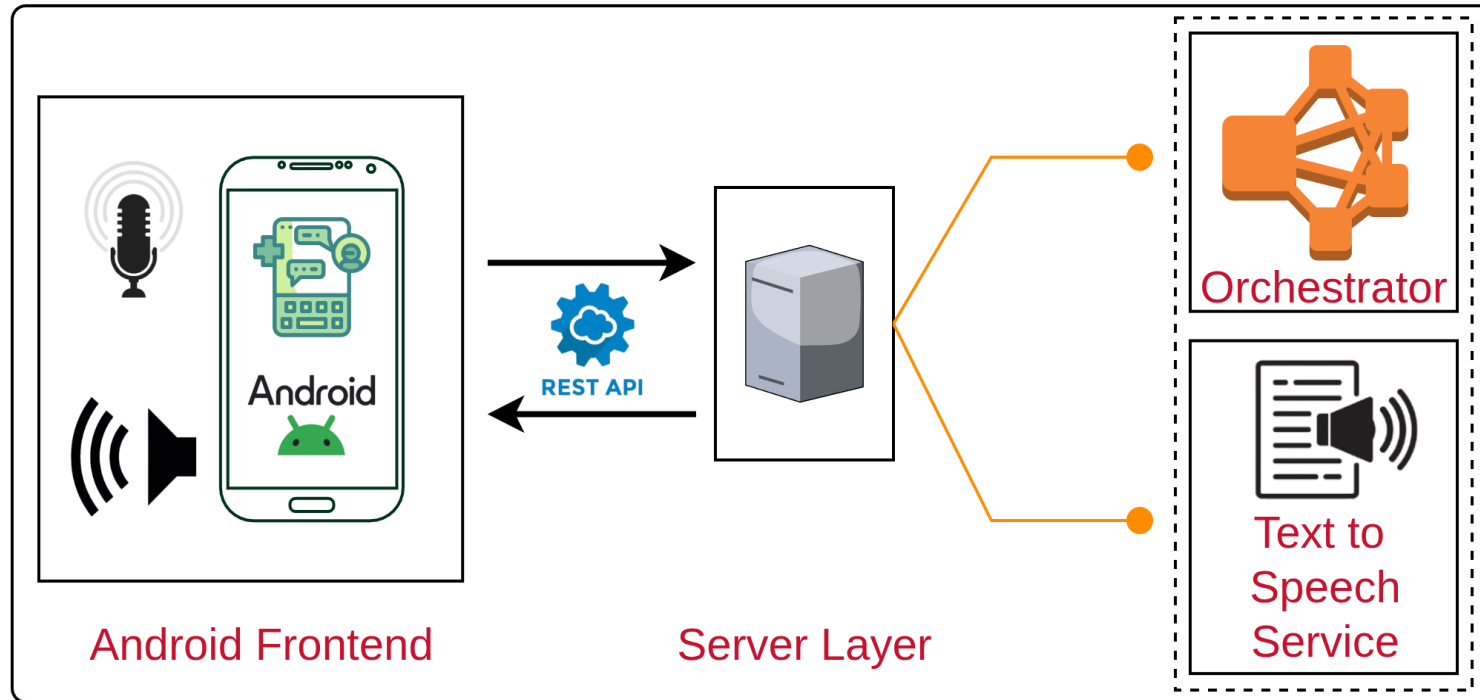
Probable diagnosis of **Dyspepsia** and organ of origin **Stomach** and the answer nodes highlighted for each of the 12 question



Patient Questionnaire Workflow



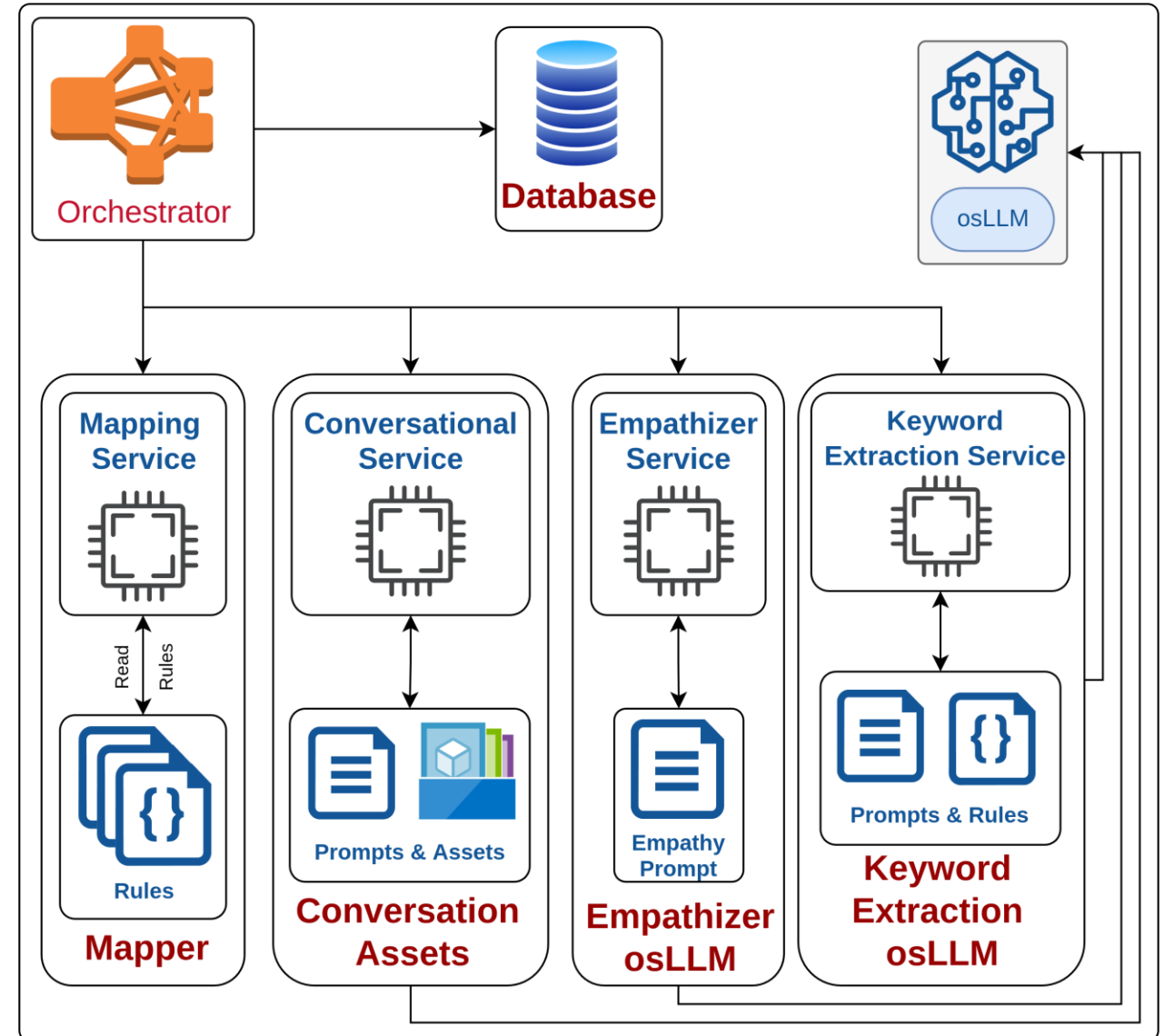
Step-by-step workflow of the patient's interaction through the questionnaire



The patient interacts with the questionnaire via the **Android application**. This app communicates with the backend **orchestrator** via a REST API through a server layer.

Android System Design

- **Orchestrator**: Coordinates calls to the specialized services and manage state.
- **Conversational Service**: Retrieving the questions text and any associated data (such as images)
- **Empathizer Service**: Rephrase the raw question into a more natural, empathic, and conversational format using **osLLMs**.
- **Keyword Extraction Service**: Extracts the relevant keywords from the patient's answer using **osLLMs**.
- **Mapper Service**: Accumulate and maps keywords from user responses to produce a ranked list of potential diagnoses.
- **Database Module**: A simple file-based store used for saving and retrieving session state.



Evaluation & System Benchmark

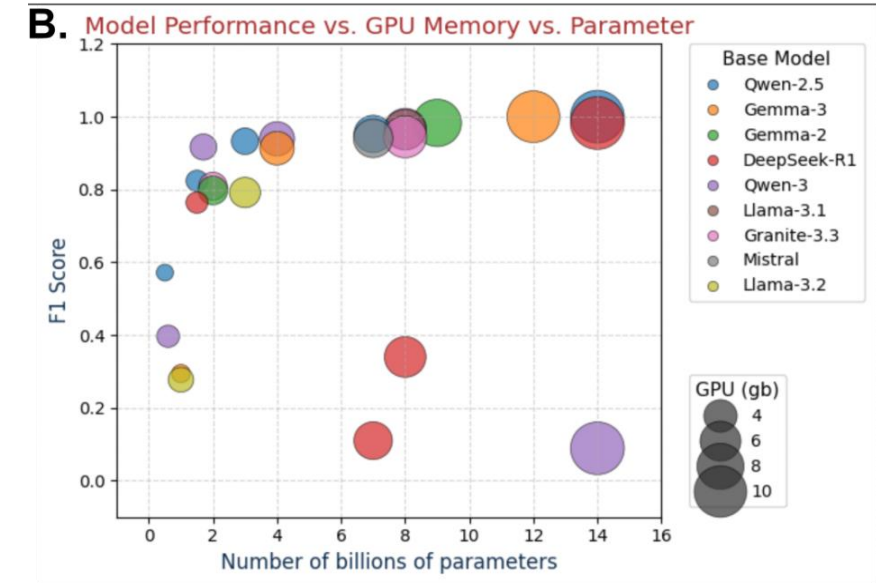
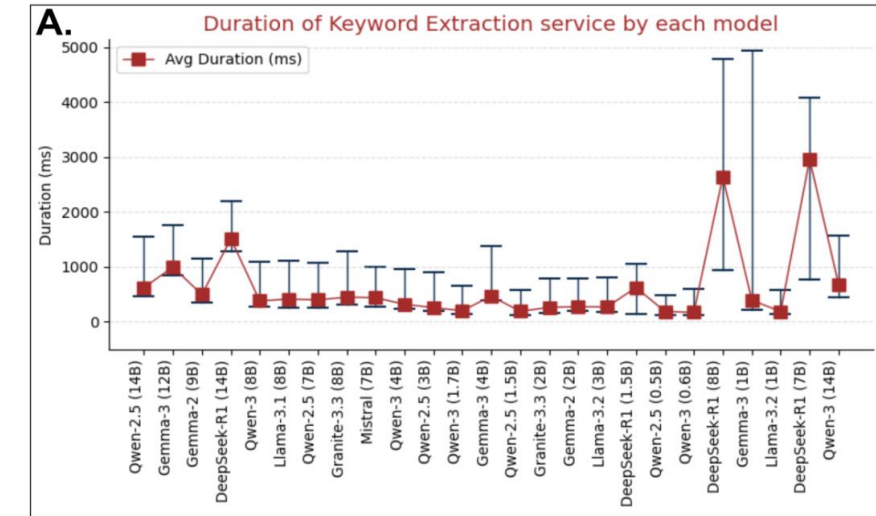
Keyword Extraction Service:

- Precision
- Recall
- F1 Score
- Latency (ms)
- GPU Memory (gb)

We selected
Gemma-3 (12B)
 model for keyword
 extraction task

Model Name	Parameter	F1	Precision	Recall	GPU
Qwen-2.5	14.0 B	1.00	1.00	1.00	10.45 GB
	7.0 B	0.95	0.98	0.93	5.28 GB
	3.0 B	0.93	0.99	0.88	2.64 GB
	1.5 B	0.82	0.94	0.73	1.69 GB
	0.5 B	0.57	0.90	0.42	1.02 GB
Gemma-3	12.0 B	1.00	1.00	1.00	10.08 GB
	4.0 B	0.91	0.91	0.92	4.25 GB
	1.0 B	0.29	0.94	0.17	1.21 GB
Gemma-2	9.0 B	0.98	0.97	1.00	8.46 GB
	2.0 B	0.80	0.91	0.71	3.08 GB
DeepSeek-R1	14.0 B	0.98	0.99	0.98	10.53 GB
	8.0 B	0.34	0.90	0.21	6.23 GB
	7.0 B	0.11	1.00	0.06	5.35 GB
	1.5 B	0.76	0.95	0.64	1.71 GB
Qwen-3	8.0 B	0.97	0.95	0.98	6.51 GB
	4.0 B	0.94	0.99	0.90	4.4 GB
	1.7 B	0.92	0.94	0.90	2.54 GB
	0.6 B	0.40	0.77	0.27	1.84 GB
	14.0 B	0.09	1.00	0.05	10.47 GB
Llama-3.1	8.0 B	0.96	0.98	0.95	6.15 GB
Llama-3.2	3.0 B	0.79	0.98	0.66	3.45 GB
	1.0 B	0.28	0.93	0.16	2.3 GB
Granite-3.3	8.0 B	0.94	1.00	0.90	6.71 GB
	2.0 B	0.81	0.78	0.84	2.9 GB
Mistral	7.0 B	0.94	0.96	0.92	5.61 GB

Qwen-2.5 (14B) and **Gemma-3 (12B)** outperform other
 osLLMs for the keyword extraction task

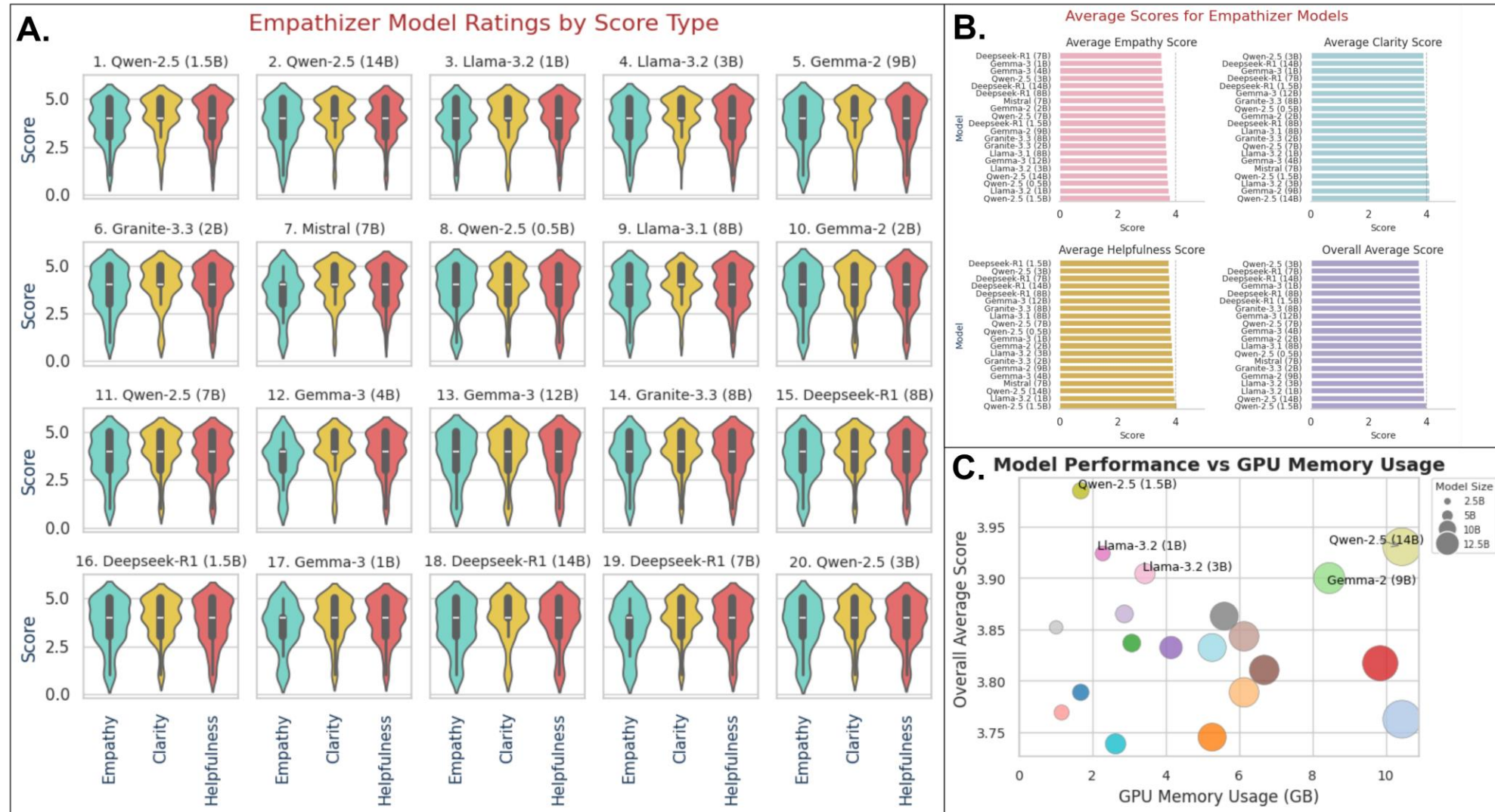


Evaluation & System Benchmark

Subjective Evaluation. All questions were rephrased into empathic question using osLLMs. Volunteers rated each question on a scale of 1-5 for the following criteria:

- Empathy
- Clarity
- Helpfulness

We selected **Qwen 2.5 (1.5B)** model for empathy task



A. Violin plot of the **scores** for each question **B.** Box plot of the **average scores** for each question **C.** **Model performance** vs **GPU memory of** osLLMs

Evaluation & System Benchmark

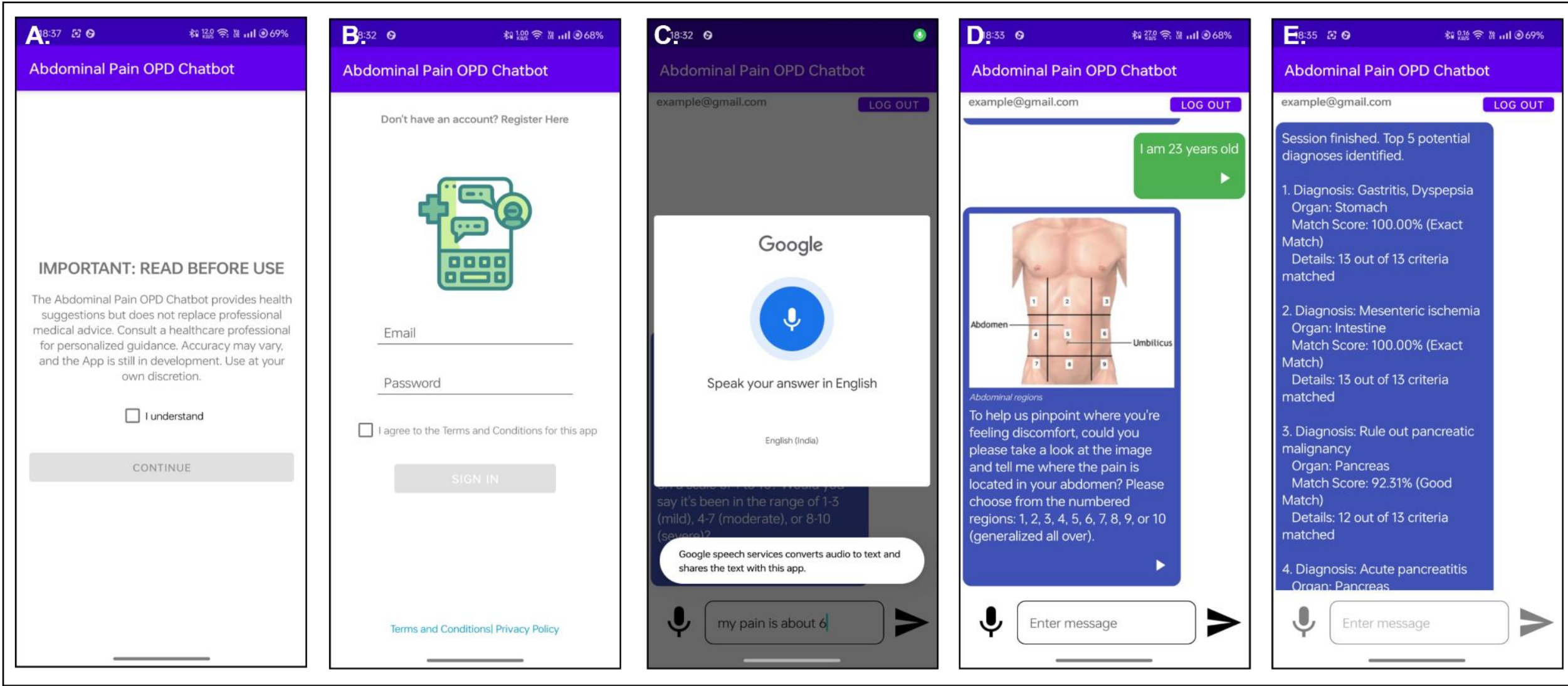
20 simulated patient interactions received from the physicians. They were evaluated in the mHealth app. The open-source large language models were used to generate the possible diagnosis without providing them the rules.

Our application outperformed existing open-source models with average top 5 accuracy of **85%** and average mean reciprocal rank (MMR) of **0.58** using Qwen-2.5 (1.5B) for empathy task and Gemma-3 (12B) for keyword extraction task.

Model Name	# Parameter	Avg Top-5 Acc	Avg MRR	GPU
App	-	85.00%	0.5808	-
Qwen2.5	14b	70.00%	0.4450	10.71 GB
	7b	45.00%	0.3167	5.41 GB
	3b	40.00%	0.1875	2.71 GB
	1.5b	35.00%	0.1617	1.73 GB
	0.5b	11.11%	0.0324	1.04 GB
Mistral	7b	60.00%	0.3433	5.75 GB
Gemma3	12b	60.00%	0.3267	10.33 GB
	4b	50.00%	0.2867	4.35 GB
	1b	11.11%	0.0741	1.24 GB
Deepseek-R1	14b	60.00%	0.2711	10.79 GB
	7b	15.00%	0.0458	5.46 GB
	8b	0.00%	0.0000	6.30 GB
	1.5b	0.00%	0.0000	1.73 GB
Gemma2	9b	50.00%	0.3183	8.67 GB
	2b	20.00%	0.0975	3.16 GB
Qwen3	1.7b	50.00%	0.1000	2.60 GB
	8b	45.00%	0.3375	6.67 GB
	4b	30.00%	0.1850	4.51 GB
	0.6b	15.00%	0.1125	1.89 GB
	14b	0.00%	0.0000	10.72 GB
Granite3.3	8b	45.00%	0.3392	6.87 GB
	2b	20.00%	0.1417	2.97 GB
Llama3.1	8b	45.00%	0.2642	6.30 GB
Llama3.2	3b	45.00%	0.1908	3.54 GB
	1b	0.00%	0.0000	2.35 GB

System evaluation results for osLLMs. The "App" is the mHealth application

Interface



Acknowledgement



Dr Anurag
Agrawal



Dr Rintu Kutum



Dr Govind K
Makharia

Clinical Scenarios



Dr Ayush
Agarwal



Dr Hara Prasad
Mishra



**Koita Centre for
Digital Health
at Ashoka**



Empathy Subjective Evaluation

- Bableen Kaur
- Sagrika Toor
- Siddhant Poudyal
- Himani Balutia
- Vinita Sharma
- Ankit Kumar
- Siddhartha Singh
- Shreyansh Priyadarshi
- Santosh Adhkari
- Tanisha Kohli