PRISM: Physician Rules Integrated with Small large language Models for probable diagnoses associated with Abdominal Pain

Gautam Ahuja[1,2,3,7,#] **Ayush Agarwal**[4,#] **Hara Prasad Mishra**[2,5,7] **Samagra Agrawal**[4]
Bableen Kaur[2,7] **Siddhant Poudyal**[2,7] **Himani Balutia**[2,7] **Sagarika**[2,7,10] **Sanjana Ahuja**[2,7]
Partha Pratim Das[1] **Ramesh Jain**[1,8] **Partha Pratim Chakrabarti**[9]
Anurag Agrawal[2,6,10†] **Govind K Makharia**[4†] **Rintu Kutum**[1,2,3,6,7†]

[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]Department of Gastroenterology & Human Nutrition, AIIMS, New Delhi, India
[5]Department of Pharmacology, University College of Medical Sciences, New Delhi, India
[6]Trivedi School of Biosciences, Ashoka University, India
[7]Augmented Health Systems Laboratory, Ashoka University, India
[8]UCI Institute for Future Health & University of California, Irvine, USA
[9]Indian Institute of Technology Kharagpur, India
[10]Department of Biology, Ashoka University, India

[#]Equal contributor
[†]Correspondence: anurag.agrawal@ashoka.edu.in,
govindmakharia@gmail.com, rintu.kutum@ashoka.edu.in,
rintu.kutum@augmented-health-systems.org

Abstract

Abdominal pain is a significant diagnostic challenge in high-volume, resource-constrained Outpatient Departments (OPDs) where it contributes to increased physician burden and potential diagnostic delays. The end-to-end workflow of Large Language Models (LLMs) offers conversational flexibility but lacks the reliability and transparency required for high-stakes clinical diagnoses (source). At the same time, traditional rule-based systems are transparent but rigid. To address this, we introduce **PRISM** (**P**hysician **R**ules **I**ntegrated with **S**mall **M**odels), a hybrid conversational system designed to augment within clinical workflows for diagnosing abdominal pain. **PRISM** utilizes a physician-guided rule-based engine for diagnostic reasoning, along with a composite of small open-source large language models (small-LLM) for patient-facing conversational (empathic) and targeted non-diagnostic tasks (relevant keyword extractions with/without UMLS). PRISM design focused and less autonomous microservices-based architecture ensures both clinical robustness and a user-centric design. PRISM achieves top-5 accuracy of $80\% - 100\%$ and Mean Reciprocal Rank (MRR) of $0.596 - 0.603$ on physician-curated 322 simulated patient question/answer pairs, outperforming the best end-to-end small-LLM with top-5 accuracy of 65% and an MMR of 0.422. Comprehensive benchmarking of empathic, clarity, and helpfulness capabilities of the small-LLMs to select the best small-LLM to be integrated with PRISM's empathy service. Likewise, benchmarking of small-LLMs was performed for the clinical keywords/terms extraction service of PRISM, using a synthetic patient Q/A dataset (SynD1) and a physician-curated simulated Q/A dataset (SimD2).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Additionally, clinical keywords/terms were extended using curated, standardized and harmonized Unified Medical Language System (UMLS) terms to evaluate the performance gain in PRISM in comparison to end-to-end small-LLMs. We introduced a multistep keyword-extraction approach to enhance clinical term extraction that leverages curated UMLS to provide focused, contextually relevant knowledge to small-LLMs, improving the clinical keyword extraction service. PRISM offers a structured, physician-in-the-loop, and resource-efficient blueprint for deploying practical generative AI applications in real-world clinical workflows.

1 Introduction

Outpatient departments (OPDs) serve as the primary point of contact for a majority of patients within a healthcare system. At All India Institute of Medical Sciences (AIIMS), New Delhi, physicians managed 0.22 million outpatient cases in 2023-2024 at the main hospital, and 0.11 million cases alone handled by the routine Department of Gastroenterology and Human Nutrition OPD, Monday through Friday, with $31,679$ new patients and $72,865$ follow-up cases. Abdominal pain is one of the most common and diagnostically challenging chief complaints encountered in outpatient settings. The differential diagnosis can span numerous organ systems and ranges from benign to life-threatening conditions, often requiring multiple diagnostic modalities [1, 2] to be considered before concluding the final diagnosis. Within this high-throughput environment, this can lead to diagnostic delays, a reduction in quality physician-patient interaction time, and increased physician burden. This highlights opportunities for digital health and GenAI technologies that can augment clinical capacity without compromising diagnostic rigor.

This paper introduces PRISM, a hybrid conversational system designed to integrate the strengths of deterministic physician-guided clinical logic (rules) with clinical vocabularies/knowledge bases and generative AI. The physician-guided deterministic rule-based approach exclusively handles the diagnostic capabilities. This ensures clinical validity, safety, and complete transparency. Natural language processing, such as targeted non-diagnostic tasks and patient-facing interactions, is managed by two small-LLMs. These small-LLMs are specialized for two distinct tasks: generating empathetic conversational questions and extracting structured clinical keywords from patients' natural language responses. This approach proposes a more focused, less automated, and transparent method for integrating generative AI into clinical workflows.

Our primary contributions are:

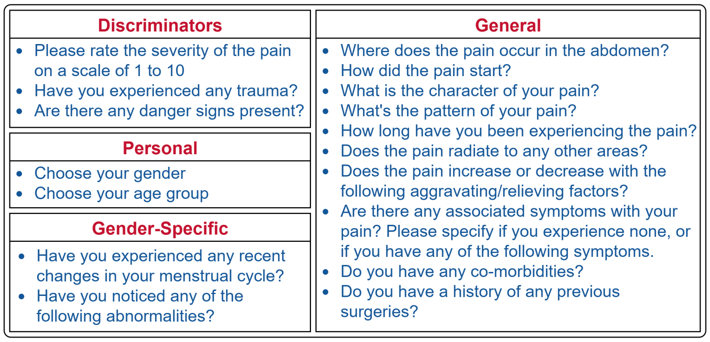
1. A hybrid conversational system - PRISM: we design and implement a microservices-based system that integrates a deterministic rule-based core with small-LLMs for a comprehensive diagnostic support tool for resource-constrained settings.
2. Targeted use of small-LLMs: we present a systematic evaluation of small-LLMs ($<$ 14B parameters) for specialized clinical subtasks, providing a benchmark for selecting resource-efficient models for real-world deployment.
3. Knowledge-enhanced extraction: we introduce a multistep extraction methodology for improving clinical keyword extraction by dynamically selecting user answer-specific knowledge bases using the curated, standardized and harmonized Unified Medical Language System (UMLS) terms [3].
4. A comprehensive benchmarking of small-LLMs: we benchmarked the GenAI-based tasks of PRISM using 28 small-LLMs for the keyword extraction service and 20 small-LLMs for the empathizer service. We also provide empirical evidence, showing the deterministic rule-based engine outperforming end-to-end LLM approaches

2 Methods

The **PRISM** is designed as a modular system that augments the initial physician-patient interaction for abdominal pain evaluation (APPENDIX A Figure 4A). The system conducts a conversational clinical assessment for abdominal pain based on a questionnaire to provide a list of probable diagnoses and the organ of origin.

2.1 Development of a physician-guided rule-based questionnaire for probable abdominal pain diagnosis

The questionnaire for probable abdominal pain diagnosis was developed and validated by three expert Gastroenterologists at the All India Institute of Medical Sciences (AIIMS), New Delhi. The questionnaire consists of 17 questions leading to 29 probable diagnoses with 19 organs of origin (APPENDIX L). The questions were categorized into four groups: discriminators, personal information, gender-specific, and general questions related to probable diagnoses (Figure 1). The discriminators' questions are used to filter potential emergency cases, might come to OPD accidentally. Personal information includes gender and age. gender-specific questions on irregularity in menstrual cycles. General sections cover ten questions with images/options (APPENDIX L).



Discriminators

- Please rate the severity of the pain on a scale of 1 to 10
- Have you experienced any trauma?
- Are there any danger signs present?

Personal

- Choose your gender
- Choose your age group

Gender-Specific

- Have you experienced any recent changes in your menstrual cycle?
- Have you noticed any of the following abnormalities?

General

- Where does the pain occur in the abdomen?
- How did the pain start?
- What is the character of your pain?
- What's the pattern of your pain?
- How long have you been experiencing the pain?
- Does the pain radiate to any other areas?
- Does the pain increase or decrease with the following aggravating/relieving factors?
- Are there any associated symptoms with your pain? Please specify if you experience none, or if you have any of the following symptoms.
- Do you have any co-morbidities?
- Do you have a history of any previous surgeries?

Figure 1: Physician-guided rule-based questionnaire developed for probable diagnoses associated with abdominal pain. Questionnaire categorized into four groups: discriminators, personal information, gender-specific, and general questions related to probable diagnoses

2.2 PRISM: Design and System Architecture

The PRISM frontend is an Android-based application featuring a conversational interface that supports both text and voice input and displays text and images to guide patients through the questionnaire. The frontend communicates exclusively with the backend via the orchestrator (Figure 2A). The frontend can also communicate with an independent text-to-speech service (Figure 2A), utilizing the Kokoro model (an open-weight text-to-speech model with 82 million parameters, `https://huggingface.co/hexgrad/Kokoro-82M`) [4].

PRISM backend is implemented using a microservice-based architecture, where independent services communicate via designated API endpoints such as i. conversational, ii. empathizer and clinical keyword extractor, and iii. mapper with the orchestrator (Figure 2B). The orchestrator maintains the control flow of the complete conversation session, such as conversational assets, an emphathizer, a keyword extractor, and a mapper service, which provides a list of probable diagnoses. Each session is stored in a JSON object with a universally unique identifier (UUID).

The control flow of the PRISM mHealth application is as follows:

1. The orchestrator greets and starts a new session.
2. It retrieves the next question and related assets (images, if any) from the Conversational Service, which stores the clinical protocol provided by the Physicians.
3. It sends the raw question to the Empathizer (LLM). Here, the question is paraphrased into natural, empathic, conversational language.
4. The patient's response is sent back to the orchestrator as free text.
5. The orchestrator passes the user's response, the original question, and the list of possible answer options to the Keyword Extractor.
6. The Keyword Extractor identifies the relevant clinical keywords from the response using the LLM and returns them to the orchestrator, which stores them.
7. This loop continues until all required questions have been answered.

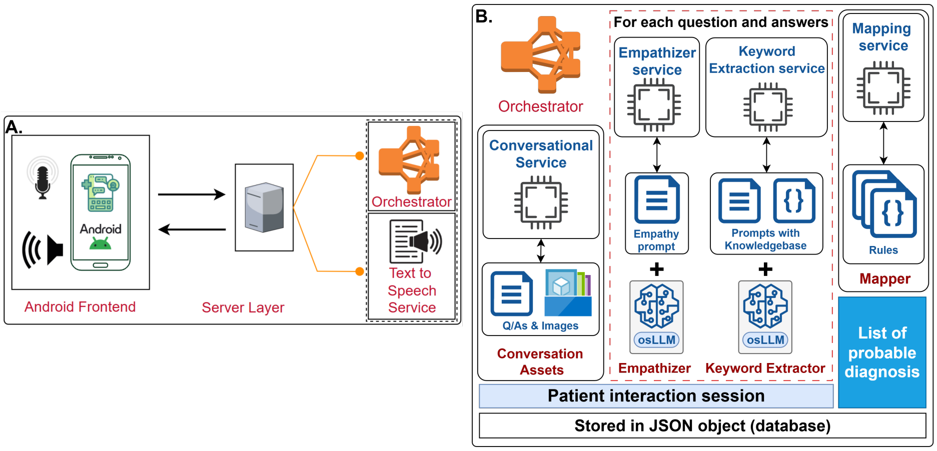


Figure 2: PRISM System Overview. (A) High-level architecture showing the Android frontend, server layer, and orchestrator. (B) Detailed backend architecture with conversational, empathizer, keyword extraction, and mapping services.

8. Finally, the orchestrator sends the complete set of extracted clinical keywords to the Mapper Service, which returns the final ranked list of probable diagnoses.

2.3 LLM for benchmarking empathizer service and clinical keyword term extraction

We have restricted our benchmarking only to LLM with parameters up to 14 billion. Models were downloaded from Ollama and used the Ollama framework to build the LLM services [5]. A complete list of model versions is in Appendix D. For the empathizer service, we benchmarked models with and without reasoning capabilities such as DeepSeek-R1 [6], Gemma-2 [7], Gemma-3 [8], Granite-3.3 [9], Llama 3.1 [10], Llama 3.2 [11], Mistral [12], and Qwen 2.5 [13]. For the keyword extraction task, Gemma3n [14], Phi-4 Reasoning [15], and Qwen-3 [16], in addition to the previously mentioned models, were benchmarked.

2.3.1 The Empathizer Service

The role of the Empathizer is to make the questionnaire protocol more empathic towards the patient while also maintaining clinical meaning. It takes the defined clinical question from the questionnaire (e.g., "What is the character of your pain?") along with its options (e.g., Burning Pain, Stabbing Pain, Pin Pricking Pain, Constricting Pain, Throbbing Pain, Dull aching/non-specific Pain) and rephrases it into a more natural and empathetic prompt (e.g., "To help me understand what you're feeling, could you describe the character of your pain? Is it more like a burning pain, a stabbing pain, a pin-pricking pain, a constricting pain, a throbbing pain, or perhaps a dull aching or non-specific pain?").

2.3.2 The Keyword Extractor Service

The Keyword Extractor is one of the most critical components of the PRISM hCS. It is responsible for parsing unstructured patient responses and extracting the specific clinical keywords that correspond to the predefined answer options in the questionnaire. This allows the system to parse a patient's narrative (e.g., “It's mostly a constant dull ache, but occasionally there are sharp, stabbing pains that come and go.”) into a structured format (`["Stabbing Pain", "Dull aching/non-specific Pain"]`) corresponding to the predefined answer options that the deterministic Mapper Service can process.

We designed and benchmarked the LLMs :

1. **Baseline**: The baseline small-LLM model is given the question, the user's answer, and a list of keywords, and is instructed to extract the ones mentioned from the user's answer.
2. **Baseline with UMLS terms**: Here, we incorporated additional clinical terms corresponding to each keyword, which were curated using the UMLS terms (APPENDIX **??**), which

encompass clinical vocabularies and terminologies, such as MTH, SNOMED-CT, MEDCIN, NCI, and CCPSS, among others. In this, the baseline prompt context is enriched with a subset of the curated UMLS terms relevant to the question as additional synonyms in the prompt. The goal is to equip the model with a comprehensive medical vocabulary to enhance recognition and extract the relevant terms.

3. **Baseline with targeted knowledge injection**: Our proposed two-stage method is designed for targeted knowledge injection.

   (a) **Step 1 (UMLS Knowledge Filtering)**: The LLM first analyzes the user's answer to identify which of the high-level answer options are most relevant. For example, if a user says "the pain is like a fire in my stomach," this step would identify "Burning Pain" as the most relevant keyword category from the full UMLS knowledge base and collate a list of all UMLS synonyms for this keyword category.

   (b) **Step 2 (Extraction)**: A second prompt is constructed for the final extraction. This prompt includes the question, the user's answer, and only the UMLS synonyms for the filtered, relevant keywords identified in Step 1. This provides the LLM with focused, contextually appropriate knowledge, reducing the context from irrelevant synonyms.

The extracted term is further utilized by the Mapper service.

2.4 The Mapper Service

PRISM's clinical reliability is the mapping service, which forms the deterministic core of the workflow. This maps the structured set of patients' extracted clinical keywords to 29 probable diagnoses, each linked to one of 19 potential organs of origin. Upon receiving a complete set of extracted keywords from the orchestrator, the mapper calculates a quantitative match score for each diagnosis based on the number of criteria met defined in the rule set. The match scoring allows for partial matches in cases of incomplete or ambiguous extracted keywords. A qualitative assessment score is also computed based on the criteria in Table 1. The diagnoses are then ranked based on their match scores, and there can be multiple matches within each rank. Two types of ranking methods were used to select the top 5 probable diagnoses: (1) M1 version assigns the rank based on the match score order, and (2) M2 version improves upon the M1 version by collating all diagnoses that achieve the same score into a single ranked set (Appendix Figure 13). For example, if three diagnoses have the highest score, the output is `([D1, D2, D3], rank 1)`. The final output is a ranked list of the top 5 potential diagnoses, each with its organ of origin, quantitative score, and qualitative assessment of the match quality.

Table 1: Assessment criteria of the mapping service

Score Range	Match Quality	Description
1	Exact Match	Perfect match (100%)
0.8-0.99	Good Match	Matched 80% or more
0.6-0.79	Fair Match	Matched 60% or more
0.4-0.59	Partial Match	Matched 40% or more
0.2-0.39	Weak Match	Matched 20% or more
>0-0.19	Poor Match	Matched at least one question
0	No Match	No match at all

2.5 Human rating of paraphrased emphatic response data of LLMs

A corpus of paraphrased emphatic response data corresponding to 17 questions was generated using the 8 LLMs with varying model parameters (a total of 20). Rating was done in a random fashion without the model information by 9 users. For the ease of rating, a Flask application was developed to assess the empathy, clarity, and helpfulness in the paraphrasing of the questions (APPENDIX I).

2.6 Patient interaction data for evaluation of the keyword extraction task and mapping service

2.6.1 Synthetic patient interaction data (SynD1)

To evaluate the keyword extraction and mapping service, we used the question-answer pairs to generate 60 patient interaction data using Gemini 2.5 Pro (prompt in APPENDIX M). This synthetic question-answer pair consists of varying difficulty levels (easy, medium, hard), and an example is listed in Table 2. The difficulty was defined by:

1. **Easy**: Focus: Simple answers, often direct, usually 1 keyword
2. **Medium**: Slightly more complex answers, potential for 2-3 keywords
3. **Hard**: Complex answers, negation, irrelevant info, multiple keywords scattered

Table 2: Example of the difficulty of the synthetic patient interaction data

Question	Patient's Answer	Extracted Keywords	Difficulty Type
What is your age?	I am 30 years old.	`25-45`	Easy
Using the diagram (regions 1-10), where is the discomfort mainly located?	It's mostly in region 1, but sometimes I feel it in region 2 as well.	`1, 2`	Medium
Any past surgeries involving the Gall Bladder, Intestine, Kidney, or Uterus?	I had my gall bladder out many years ago, and also uterine surgery. Nothing involving the intestine or kidney.	`Gall Bladder, Uterus`	Hard

2.6.2 B. Simulated patient interaction data (SimD2)

A set of 20 simulated patient interaction data was provided by Gastroenterologists and Clinical Pharmacologists, which was used for the evaluation of the LLM in keyword extraction and mapping the top 5 probable diagnoses. The simulated data consisted of 322 question-answer pairs in total.

2.7 Evaluation metrics

We selected the following metrics for each task:

1. **Empathizer service**: Raters scored each rephrased question on a scale of 1 to 5 for Empathy, Clarity, and Helpfulness. **Empathy rating (1-5)**: How well does the rephrased question convey warmth, understanding, and respect for the patient's experience? **Clarity rating (1-5)**: How clear, simple, and easy is the rephrased question to understand for a patient without a medical background? **Helpfulness rating (1-5)**: How effectively does the question guide the patient to provide the necessary clinical information while retaining the intent of the original question?
2. **Keyword extraction service**: Performance was measured using standard classification metrics: Precision, Recall, and F1-Score.
3. **Mapping service**: The final ranked list of diagnoses was evaluated using Top-5 Accuracy (whether the correct diagnosis is within the top 5 suggestions) and Mean Reciprocal Rank (MRR), which measures the quality of the ranking.
4. **System Efficiency**: GPU memory usage and response latency were also measured for all component benchmarks.

2.8 Curation of additional associated clinical keywords/terms using UMLS

The questions and answers/options dataset were used for expanding the 76 clinical keywords/terms using the UMLS knowledgebase. Curation was performed in two stages: firstly, with a custom Python script using the UMLS API (`https://uts-ws.nlm.nih.gov/rest/search/current?apiKey=`

`API_KEY&string=abdominal+pain&pageSize=500`), secondly, cross-validating the outputs with the UMLS browser (`https://uts.nlm.nih.gov/uts/umls/searchResults?searchString=abdominal%20pain`), and only 182 relevant clinical terms were retained, covering 13 biomedical vocabularies, terminologies, and ontologies; and irrelevant terms were removed. Secondly, we expanded the clinical terms with their designated preferred name, designated synonym, nursing indicator, full form of descriptor, lower level term, entry term, metathesaurus preferred name, etc, covering 63 biomedical vocabularies/terminologies/ ontologies using the CUI (Concept Unique Identifier) based UMLS API `https://uts-ws.nlm.nih.gov/rest/content/current/CUI/C0000737/atoms?apiKey=API_KEY`, where C0000737 corresponds to abdominal pain. The expanded clinical terms correspond to a total of 3800 standardized terms. Harmonization of the terms, such as removing qualified values, lowercasing, and special character removal, led to 1164 curated clinical terms corresponding to 76 original clinical keywords/terms.

2.9 Android-based mobile health (mHealth) app development

An Android-based mHealth app was developed as a proof of concept in an intranet setting using the top-performing small LLM (Qwen-2.5-1.5B for emphatizer service and Qwen-2.5-14B for the keyword extraction task) without the additional knowledge injection.

3 Experiments, evaluation and results

We conducted a series of experiments for the services in PRISM, two that utilize small-LLMs, namely, Empathizer and Keyword Extractor services, and ranking methods for the mapper service.

3.1 Qwen-2.5 with 1.5 billion parameters ranked 1 in empathy and helpfulness through human rating

To evaluate the paraphrasing of the questions by the small LLMs, the rating was performed in three categories: empathy, clarity, and helpfulness by the human raters (rating system is described in Methods). In empathy and helpfulness, Qwen-2.5 1.5B ranked 1 with 3.82 ± 0.09 and 4.05 ± 0.09 human rating scores, respectively. In the clarity, Qwen-2.5 14B ranked 1 with 4.12 ± 0.08. Overall, based on the three categories, Qwen-2.5 1.5B ranked 1 with an average rating of 3.99 ± 0.08. The top 5 are shown in the Table 3

Table 3: Benchmark of small-LLMs for Empathizer Service (Top 5 shown)

Base model	# Para.(B)	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

The complete results of the evaluation are in APPENDIX Table 9.

3.2 Evaluation of the keyword extraction service from the patient interaction data and mapping service

To evaluate the small-LLM's capability in the clinical keyword extraction task, two datasets, namely, the synthetic patient interaction (PI) dataset (synD1, N=60 Q/A-pairs), validated by general physicians, and a simulated PI dataset (simD2, N=322 Q/A-pairs) by gastroenterologists and general physicians/clinical pharmacologists, were used. In the SynD2, Qwen-2.5 (14B) and Gemma-3 (12B) outperformed with a F1 score of 1.0 and 0.91 in the baseline and baseline with UMLS terms settings, respectively, as described in the Method section. In the baseline with the targeted knowledge injection setting, we observed Phi-4 (14B) reasoning model and Gemma-3 (12B) scored an F1 score of 1.0, outperforming other LLMs. 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). Based on the overall performance in SynD1 and SimD2, Qwen-2.5 (14B) was considered for the keyword extraction service. The extracted clinical terms are further linked with the diagnosis and the clinical terms dictionary to find the top five probable diagnoses associated with abdominal pain using the mapping service of PRISM. The objective of developing PRISM was to provide a list of probable diagnoses associated with abdominal pain from the 29 diagnoses provided by the Gastroenterologists from AIIMS, New Delhi, India. It provides both exact and partial mapped diagnoses with the clinical keywords/terms.

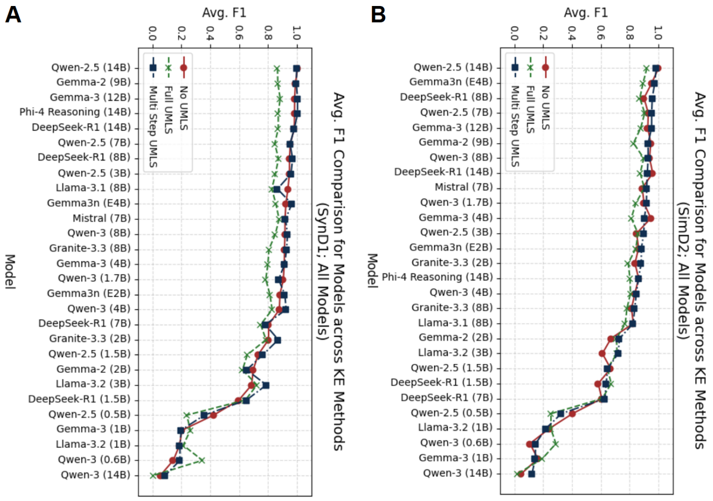


Figure 3: Benchmarking of across small-LLMs with average F1 (Avg. F1) scores for (A) SynD1 and (B) SimD2 dataset

3.3 PRSIM outperforms the end-to-end small open source LLM workflow in probable diagnoses

To evaluate the performance of PRISM and the end-to-end small-LLM workflow, we used SimD2, and performance was quantified using two methods: **M1**. simple ranking, which is based on the number of clinical terms (CT) mapping corresponding to the diagnoses, only the top five (for example, R1: PD7 with 15 CT - extract match, R2: PD1 with 14CT, R3: PD11 with 14CT, R4: PD1 with 13CT, and R5: PD9 with 12CT). **M2**. A set ranking, which is based on the number of clinical terms associated with diagnoses, was converted into a set rank, for example, R1: PD7 with 15CT, R2: PD1, PD11 with 14CT, R3: PD9, PD29, PD15 with 13CT, R4: PD3, PD8 with 12CT, and R5: PD4, PD2 11CT.

The evaluation of PRISM was performed against a total of 11 models with varying parameter sizes, providing the physician-curated Q/A pairs and the list of 29 diagnoses, except for the rules/associated links of the clinical terms associated with these 29 diagnoses. Overall, PRISM outperforms all 11 end-to-end small-LLM workflows with a top 5 (Top5) accuracy of 0.8 and a mean reciprocal rank (MRR) of 0.57, reflecting the expected probable diagnosis by the Gastroenterologist was observed in the top 2 ranks based on the **M1 ranking method**. The PRISM's performance with the **M2 ranking method**, we observed a Top5 accuracy of 1.0 and a mean reciprocal rank (MRR) of 0.59.

In the end-to-end LLM models, Phi-4 (14B) with reasoning and Qwen-2.5 (14B) outperformed other end-to-end models with a Top5 accuracy of 0.65 (MRR=0.47) and 0.65 (MRR=0.39), respectively.

This highlights the potential of PRISM - the physician-guided rules and/or integration of rich biomedical vocabularies, with a narrowly defined task for small language models, integrated as a mobile/tablet Health conversational application (contextual), has the potential to assist in low-resource constrained hospital infrastructures.

The full table is in APPENDIX 10

Table 4: Performance of PRISM and end-to-end small-LLMs (Top 5 shown)

Model Name	**# Para. (B)**	**Avg. Top-5 Accuracy**	**Avg. MRR**
PRISM (M2 Ranking)	*	**1.00 ± 0.00**	**0.59 ± 0.08**
PRISM (M1 Ranking)	*	**0.8 ± 0.09**	**0.57 ± 0.09**
Phi4-Reasoning	**14**	**0.65 ± 0.11**	**0.47 ± 0.10**
Qwen-2.5	**14**	**0.65 ± 0.11**	**0.39 ± 0.09**
Mistral	7	0.60 ± 0.11	0.36 ± 0.09
Gemma-3	12	0.60 ± 0.11	0.31 ± 0.08
Gemma3n	E4	0.55 ± 0.11	0.35 ± 0.09

*Qwen-2.5 (14B) is used for keyword extraction and Qwen-2.5 (1.5B) for emphatic responses.

4 Limitations

Despite the promising results of PRISM, there are many shortcomings in the current work. Firstly, the synthetic dataset 1 (SynD1) contains limited Q/A pairs (N=60) from only Gemini 2.5 Pro, which is from only one distribution, which might induce bias in SynD1 and may not generalize. Toward this, future experiments will incorporate a larger corpus of Q/A pairs with varying closed LLMs with a larger parameter space. Secondly, the simulation dataset (SimD1) contains only 322 Q/A pairs, and a larger simulated Q/A dataset from multiple hospitals would help us understand the stability and the performance of the small LLMs in paraphrasing the questions regarding empathy, clarity and helpfulness; and regarding clinical keywords/terms extraction from the users' responses. Thirdly, the rating of the paraphrasing was done by only nine raters, and rating with more users with the Flask app would help us better understand the characteristics of stratified user groups with varying levels of education and exposure in social science and health science professionals and conduct a survey among patient groups. Fourthly, the PRISM mHealth App has been tested in a controlled environment. In future editions, testing the application independently with a general physician and a gastroenterologist would allow us to evaluate PRISM's performance better. Lastly, the mapping of clinical terms with negation terms, such as mapping "No" couldn't be mapped to the "None" option, introduces false negatives. We intend to improve on this in the future version of PRISM.

5 Discussion and future directions

Generative AI Models bring both opportunities and challenges while being deployed in healthcare clinical workflow, clearance of government regulatory agencies, both national and international, and in resource-constrained settings within a clinical ecosystem. PRISM was designed keeping these challenges in mind, where we utilized small open-source language models and applied them in narrow tasks such as conversational tasks, keyword extractions guided with external, well-curated clinical keywords/ terms such as UMLS. Currently, the introduction of a knowledge base has not improved the model performance, but our assumption on utilizing the knowledge base is primarily to support the small language models in wild scenarios/real-world settings.

Currently, the conversation service is layered with Google Firebase Version for speech-to-text transcription, and for text-to-speech, we have used Kokoro, an open-weight TTS model with 82 million parameters. We intend to incorporate Indic Models and other models, such as Qwen-MT, and benchmark them in well-curated transcripts and translations.

6 Conclusion

This paper introduced PRISM, a hybrid conversational AI system that combines a deterministic, physician-curated rule-based engine with specialized small open-source large language models to assist in the diagnosis of abdominal pain. We find that the current system achieves a more accurate diagnosis compared to end-to-end LLM approaches. Furthermore, our Multistep UMLS keyword extraction method shows that targeted and more contextual knowledge injection is more effective than broad data augmentation. PRISM offers an effective blueprint for developing GenAI tools in healthcare that are focused, resource-optimized, transparent, and designed for real-world clinical settings.

References

[1] Sarah L Cartwright and Mark P Knudson. Evaluation of acute abdominal pain in adults. *American family physician*, 77(7):971–978, 2008.

[2] Sarah L Gans, Margreet A Pols, Jaap Stoker, Marja A Boermeester, and Expert Steering Group. Guideline for the diagnostic pathway in patients with acute abdominal pain. *Digestive surgery*, 32(1):23–31, 2015.

[3] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[4] hexgrad. Kokoro-82m. `https://huggingface.co/hexgrad/Kokoro-82M`.

[5] ollama. Ollama. `https://ollama.com/`.

[6] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[7] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[8] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[9] IBM. ibm-granite/granite-3.3-8b-instruct. `https://huggingface.co/ibm-granite/granite-3.3-8b-instruct`.

[10] Meta. Llama 3.1. `https://ai.meta.com/blog/meta-llama-3-1/`.

[11] Meta. Llama 3.2. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`.

[12] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[13] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[14] Google. Gemma3n. `https://deepmind.google/models/gemma/gemma-3n/`.

[15] Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report, 2025.

[16] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin,

Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

A System Hardware Details

- Workstation: Dell Precision 5820 Tower
- OS: Ubuntu 22.04.3 LTS
- Memory: 512GB
- CPU: Intel(R) Xeon(R) W-2295 CPU @ 3.00GHz
- GPU: 1 x NVIDIA RTX A5000 (24GB)
- Storage: 1 x NVMe SK hynix 512GB

B Detailed individual author contribution

C Patient workflow

PRISM augments the physician-patient interaction of the abdominal pain diagnostic process. By providing a list of probable diagnoses, the hCS aims to reduce clinical burden and allow physicians to have a more targeted session during the interaction. The current framework consists of 17 questions divided into four categories: Discriminator, Personal, Gender specific, and General questions. A patient would first be greeted and asked the Discriminator questions. Any positive responses to questions regarding trauma, danger signs, or pain intensity more than 8 trigger an emergency department redirect. Regular OPD patients are asked personal questions regarding age and gender. Female patients are redirected to answer questions regarding their menstrual health, followed by General questions, while males move to answering General questions directly after answering Personal questions. After answering all questions, the patients are provided with a list of probable diagnoses and the respective organs of origin.

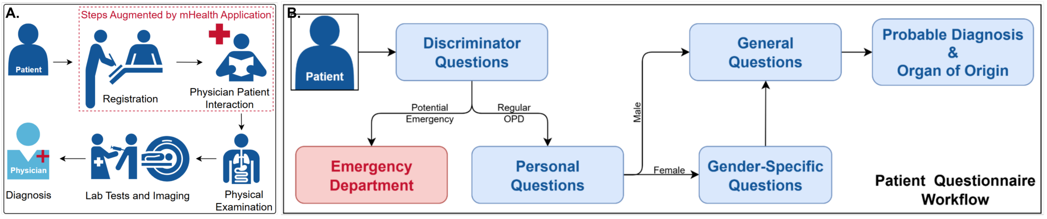


Figure 4: Patient Workflow

D Model version

We used the Ollama framework to build the LLM services, with models downloaded from the Ollama model library (`https://ollama.com/library`). The Table 5 provides details about the downloaded models, including their Ollama model names, version IDs, sizes, quantization, types, and the benchmarks for which they were used - Keyword Extractor (KE) and Empathizer.

Table 5: List and Specifications of Ollama Models Used in Benchmarks

MODEL_NAME	VERSION_ID	SIZE	QUANTIZATION	BENCHMARKED
deepseek-r1:1.5b	a42b25d8c10a	1.1 GB	Q4_K_M	KE, EMPATHIZER
deepseek-r1:14b	ea35dfe18182	9.0 GB	Q4_K_M	KE, EMPATHIZER
deepseek-r1:7b	0a8c26691023	4.7 GB	Q4_K_M	KE, EMPATHIZER
deepseek-r1:8b	28f8fd6cdc67	4.9 GB	Q4_K_M	KE, EMPATHIZER
gemma2:2b	8ccf136fdd52	1.6 GB	Q4_0	KE, EMPATHIZER
gemma2:9b	ff02c3702f32	5.4 GB	Q4_0	KE, EMPATHIZER
gemma3:12b	f4031aab637d	8.1 GB	Q4_K_M	KE, EMPATHIZER
gemma3:1b	8648f39daa8f	815 MB	Q4_K_M	KE, EMPATHIZER
gemma3:4b	a2af6cc3eb7f	3.3 GB	Q4_K_M	KE, EMPATHIZER
gemma3n:e2b	719372f8c7de	5.6 GB	Q4_K_M	KE
gemma3n:e4b	15cb39fd9394	7.5 GB	Q4_K_M	KE
granite3.3:2b	07bd1f170855	1.5 GB	Q4_K_M	KE, EMPATHIZER
granite3.3:8b	fd429f23b909	4.9 GB	Q4_K_M	KE, EMPATHIZER
llama3.1:8b	46e0c10c039e	4.9 GB	Q4_K_M	KE, EMPATHIZER
llama3.2:1b	baf6a787fdff	1.3 GB	Q8_0	KE, EMPATHIZER
llama3.2:3b	a80c4f17acd5	2.0 GB	Q4_K_M	KE, EMPATHIZER
mistral:7b	f974a74358d6	4.1 GB	Q4_0	KE, EMPATHIZER
phi4-reasoning:14b	47e2630ccbcd	11 GB	Q4_K_M	KE
qwen2.5:0.5b	a8b0c5157701	397 MB	Q4_K_M	KE, EMPATHIZER
qwen2.5:1.5b	65ec06548149	986 MB	Q4_K_M	KE, EMPATHIZER
qwen2.5:14b	7cdf5a0187d5	9.0 GB	Q4_K_M	KE, EMPATHIZER
qwen2.5:3b	357c53fb659c	1.9 GB	Q4_K_M	KE, EMPATHIZER
qwen2.5:7b	845dbda0ea48	4.7 GB	Q4_K_M	KE, EMPATHIZER
qwen3:0.6b	3bae9c93586b	522 MB	Q4_K_M	KE
qwen3:1.7b	458ce03a2187	1.4 GB	Q4_K_M	KE
qwen3:14b	bdbd181c33f2	9.3 GB	Q4_K_M	KE
qwen3:4b	a383baf4993b	2.6 GB	Q4_K_M	KE
qwen3:8b	e4b5fd7f8af0	5.2 GB	Q4_K_M	KE

E UMLS knowledgebase curation

The knowledgebase was curated using UMLS. The 76 clinical keywords/terms were searched using the UMLS API and cross-validated using the UMLS browser, resulting in the retention of 182 unique relevant terms. The Table 6 shows the distribution of the 182 terms across the four question categories. The Table 7 shows the distribution of the 182 terms across 13 clinical and biomedical vocabularies. These terms were then expanded with their designated preferred name, designated synonym, nursing indicator, full form of descriptor, etc., to a total of 3800 standardized terms using the UMLS API. The Table 8 shows the distribution of the 3800 terms across different term types in UMLS.

Table 6: Distribution of the 182 terms across the 4 question categories

Category	Count
General	140
Discriminator	25
Gender-Specific	15
Personal	2

Table 7: Distribution of the 182 terms across 13 clinical and biomedical vocabularies

Vocabulary	Counts
MTH	77
SNOMEDCT_US	40
MEDCIN	32
NCI	13
CCPSS	4
LNC	4
MDR	3
HPO	2
CHV	2
ICD10CM	2
OMIM	1
COSTAR	1
MSH	1

Table 8: Distribution of the 3800 terms across different UMLS term types

Term Type	Description	Count
PT	Designated preferred name	1276
SY	Designated synonym	990
ID	Nursing indicator	254
FN	Full form of descriptor	226
LLT	Lower Level Term	174
ET	Entry term	147
PN	Metathesaurus preferred name	76
PM	Machine permutation	62
GT	Glossary term	62
PTN	"Preferred term, natural language form"	53
HT	Hierarchical term	53
PTCS	Preferred Clinical Synopsis	50
LA	LOINC answer	44
MH	Main heading	35
IT	Index term	33
RT	"Term that is related to, but often considered non-synonymous with, the preferred term"	32
FI	Finding name	31
NP	Non-preferred term	30
DE	Descriptor	25
LPN	LOINC parts name	23
LPDN	LOINC parts display name	21
CN	LOINC official component name	18
PTGB	British preferred term	14
AD	Adjective	11
SYGB	British synonym	10
MTH_PTN	"Metathesaurus preferred term, natural language form"	9
PEP	Preferred entry term	8
PR	Name of a problem	8
MD	CCS multi-level diagnosis categories	5
SD	CCS diagnosis categories	5
SI	Name of a sign or symptom of a problem	3
AB	Abbreviation in any source vocabulary	3
HC	Hierarchical class	2
DI	Disease name	1
DC10	Diagnostic criteria for ICD10 code	1
DC9	Diagnostic criteria for ICD9 code	1
HG	High Level Group Term	1
HS		1
AT	Attribute type	1
OS	System-organ class	1

F Additional details and results (Empathizer)

The corpus of 17x20 paraphrased questions was randomly rated by 9 users on a scale of 1-5 for empathy, clarity, and helpfulness. The paraphrased question order was randomized, and the model information was hidden. The user ratings were then analyzed to find the optimal osLLM for empathic question paraphrasing. The Figure 5A shows the distribution of user rating scores across three parameters for 20 variants of 8 osLLMs. Figure 5B shows the average rating scores of different parameters across all models. We see that Qwen-2.5 (1.5B) outperformed other models for empathy and helpfulness, while Qwen 2.5 (14B) outperformed others for clarity. Figure 6 shows the overall score with respect to model GPU usage and parameter size.

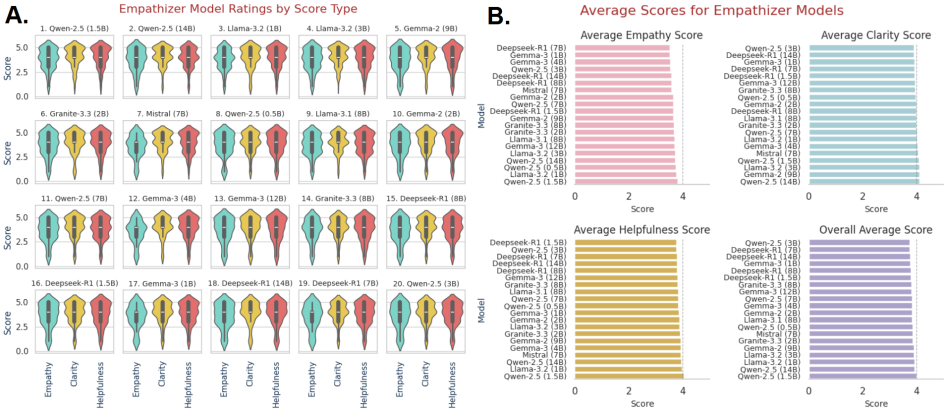


Figure 5: User rating score distribution. (A) Violin plot of user rating scores across three parameters. (B) Average rating scores of different parameters across all models

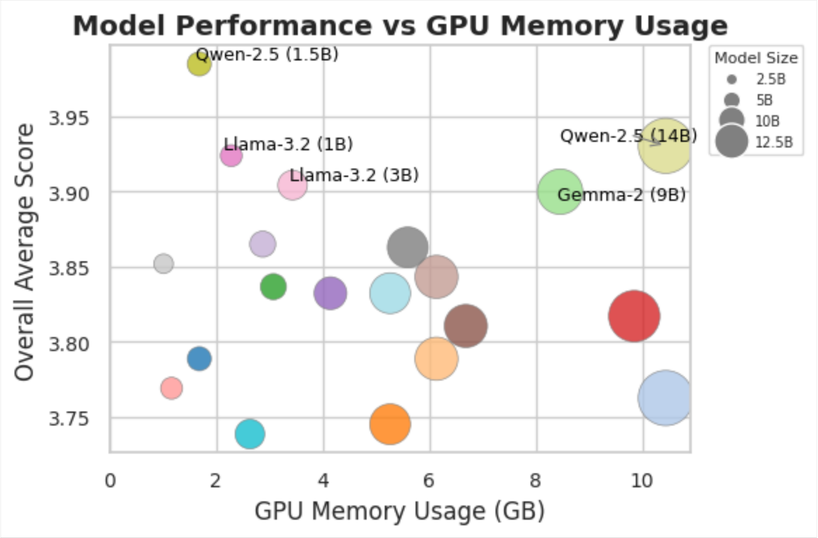


Figure 6: bubble plot

Table 9: Benchmark of small-LLMs for Empathizer Service (Full shown)

Base model	# Para.(B)	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
Granite-3.3	2	3.68 ± 0.10	4.00 ± 0.09	3.92 ± 0.09	3.87 ± 0.09
Mistral	7	3.60 ± 0.09	4.05 ± 0.08	3.94 ± 0.08	3.86 ± 0.09
Qwen-2.5	0.5	3.75 ± 0.10	3.96 ± 0.08	3.84 ± 0.09	3.85 ± 0.09
Llama-3.1	8	3.69 ± 0.09	3.99 ± 0.08	3.84 ± 0.09	3.84 ± 0.08
Gemma-2	2	3.65 ± 0.10	3.99 ± 0.09	3.88 ± 0.10	3.84 ± 0.10
Qwen-2.5	7	3.65 ± 0.10	4.01 ± 0.09	3.84 ± 0.09	3.83 ± 0.09
Gemma-3	4	3.54 ± 0.09	4.03 ± 0.08	3.93 ± 0.09	3.83 ± 0.09
Gemma-3	12	3.70 ± 0.10	3.94 ± 0.10	3.81 ± 0.10	3.82 ± 0.10
Granite-3.3	8	3.68 ± 0.10	3.94 ± 0.08	3.81 ± 0.09	3.81 ± 0.09
DeepSeek-R1	8	3.58 ± 0.10	3.99 ± 0.08	3.80 ± 0.09	3.79 ± 0.09
DeepSeek-R1	1.5	3.67 ± 0.10	3.93 ± 0.09	3.77 ± 0.10	3.79 ± 0.09
Gemma-3	1	3.53 ± 0.10	3.92 ± 0.09	3.86 ± 0.09	3.77 ± 0.09
DeepSeek-R1	14	3.58 ± 0.10	3.91 ± 0.09	3.80 ± 0.10	3.76 ± 0.10
DeepSeek-R1	7	3.52 ± 0.10	3.92 ± 0.08	3.80 ± 0.09	3.75 ± 0.09
Qwen-2.5	3	3.54 ± 0.11	3.90 ± 0.09	3.78 ± 0.09	3.74 ± 0.10

G Additional details and results (Keyword Extractor)

Synthetic patient interaction data (synD1) with 60 Q/A pairs was used to benchmark the three keyword extraction methods. The mean and median results for precision, recall, f1 and duration with bars representing the minimum and maximum values are shown in Figures 7, 8, 9

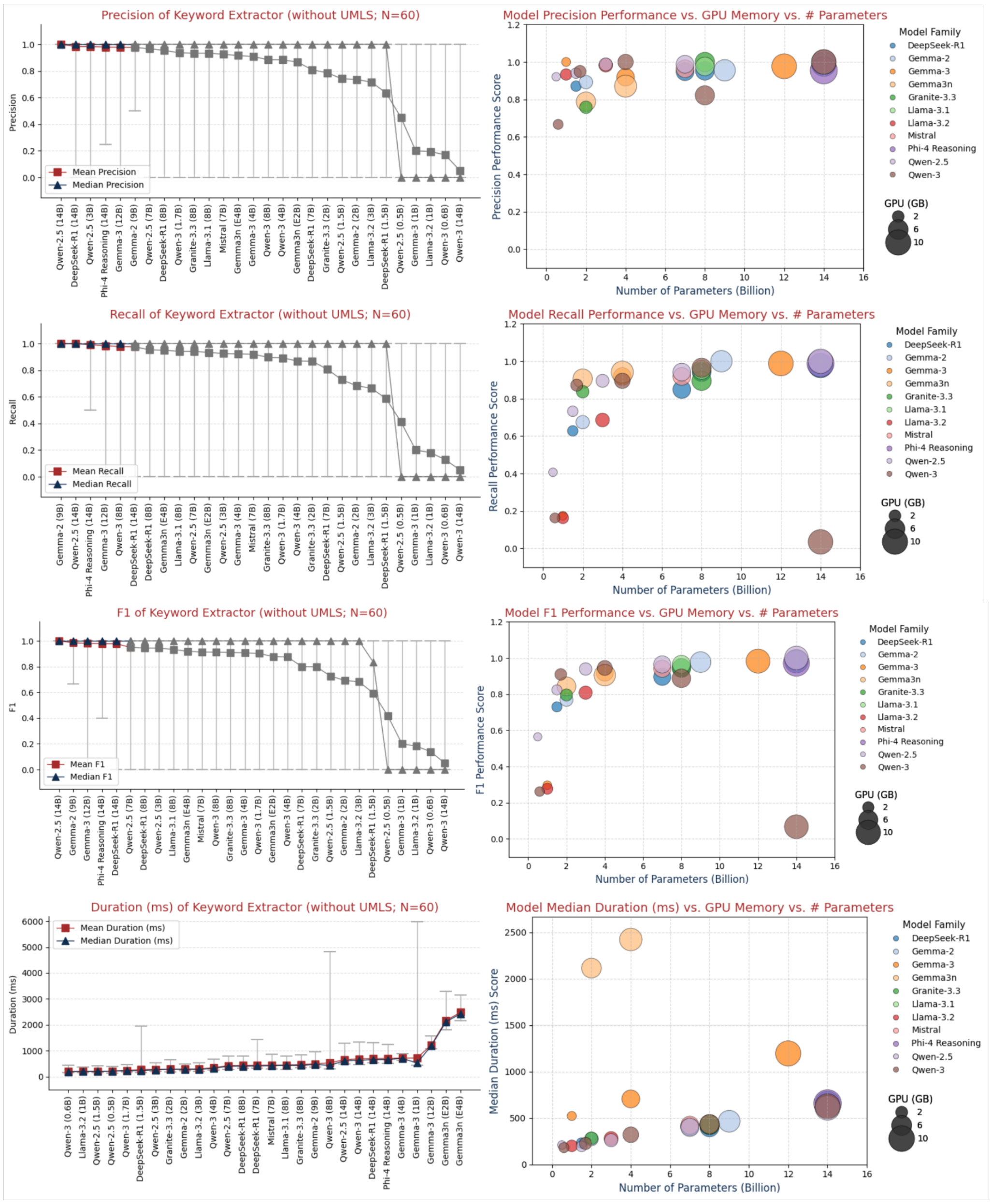


Figure 7: Baseline

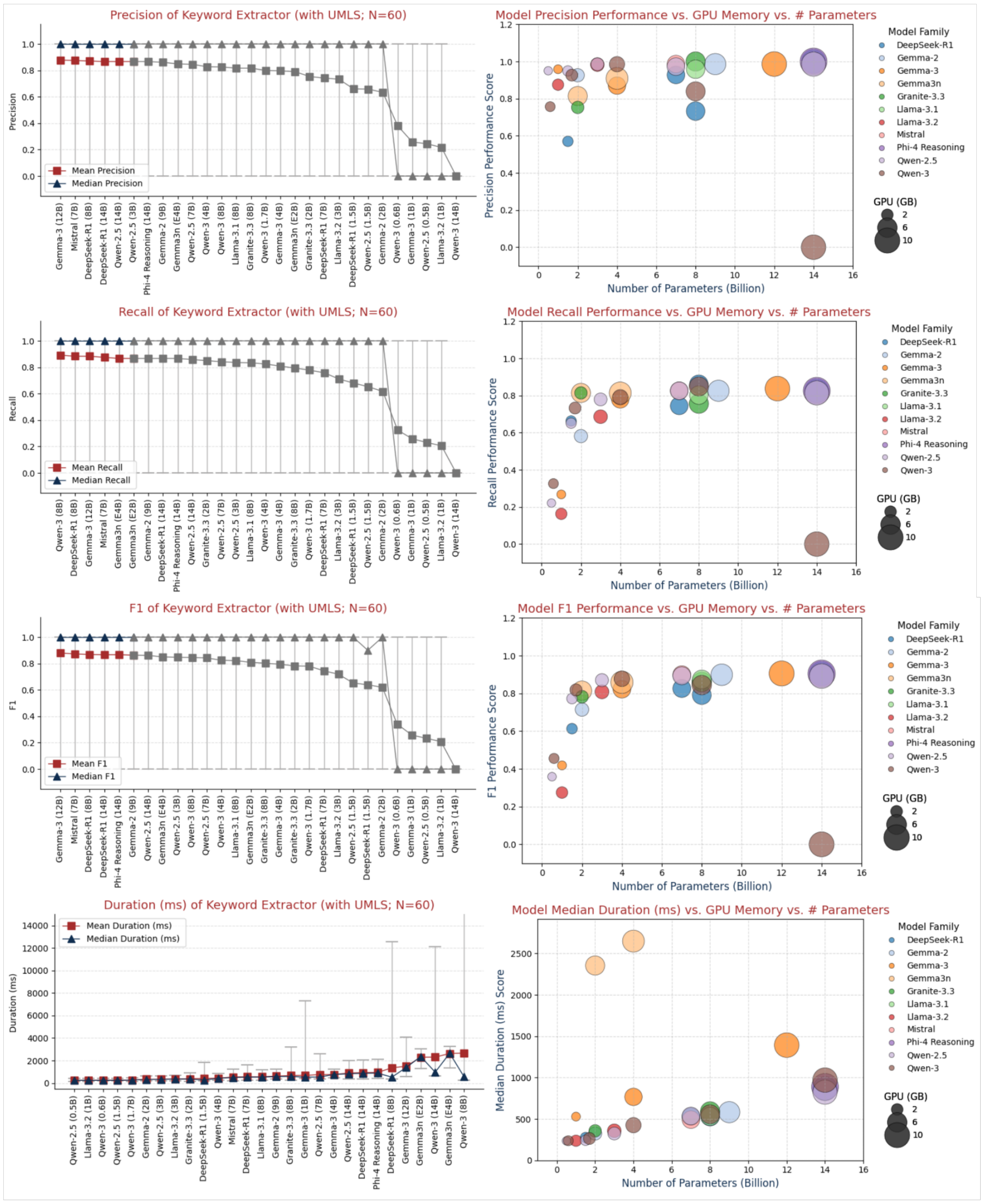


Figure 8: Baseline UMLS

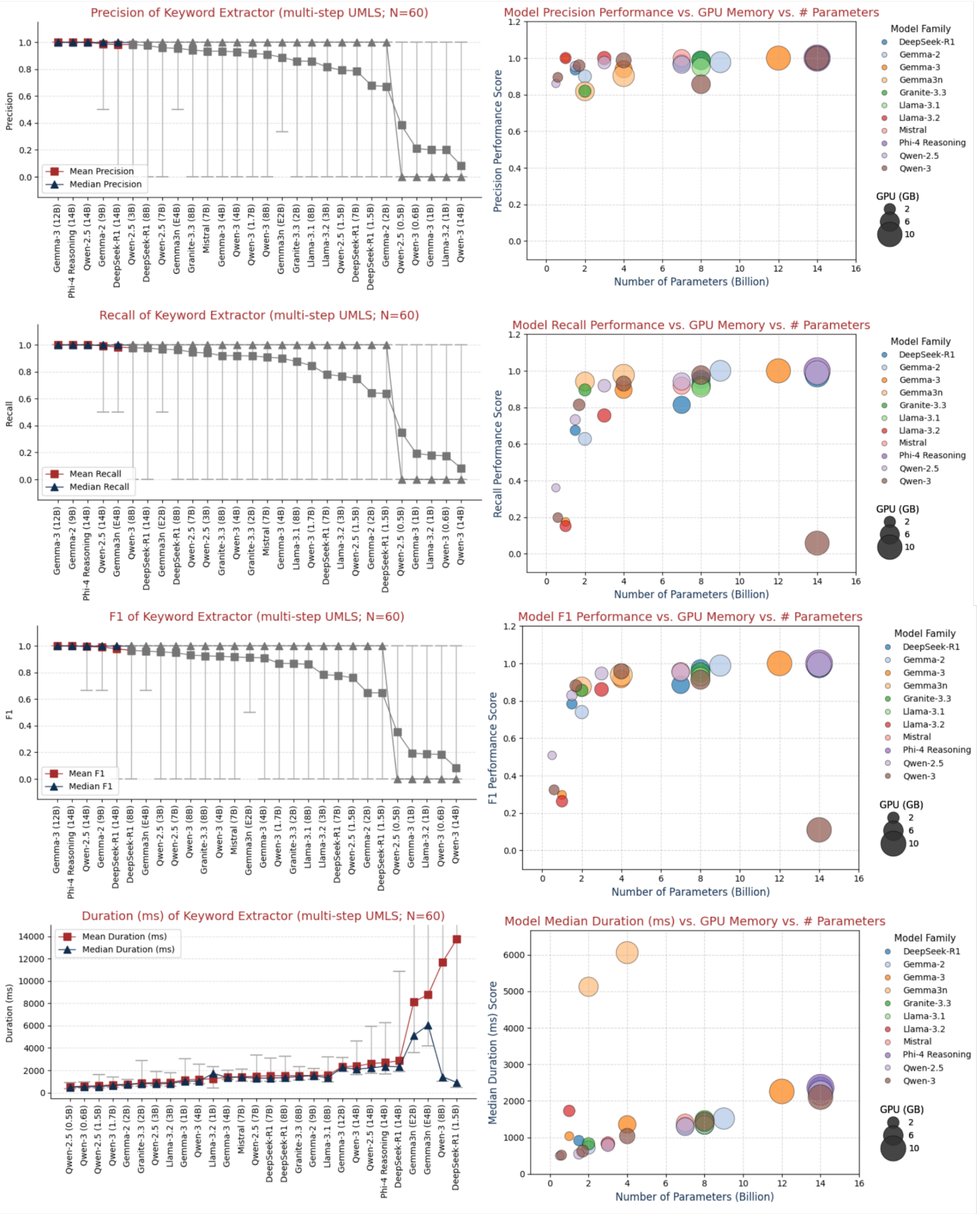


Figure 9: Multistep UMLS

Simulated patient interaction data (simD2) with 322 Q/A pairs was used to benchmark the three keyword extraction methods. The mean and median results for precision, recall, f1 and duration with bars representing the minimum and maximum values are shown in Figures 10, 11, 12

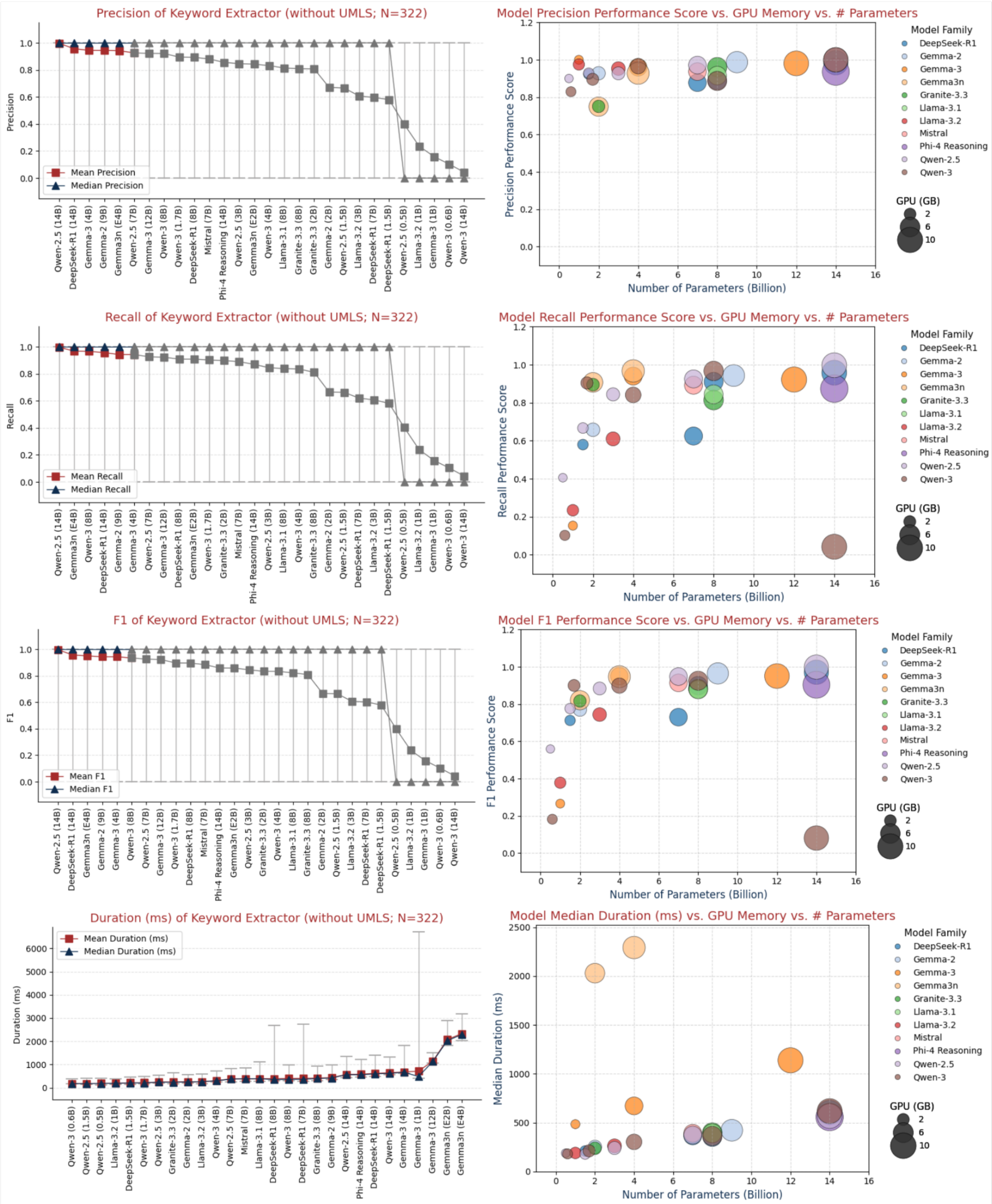


Figure 10: Baseline

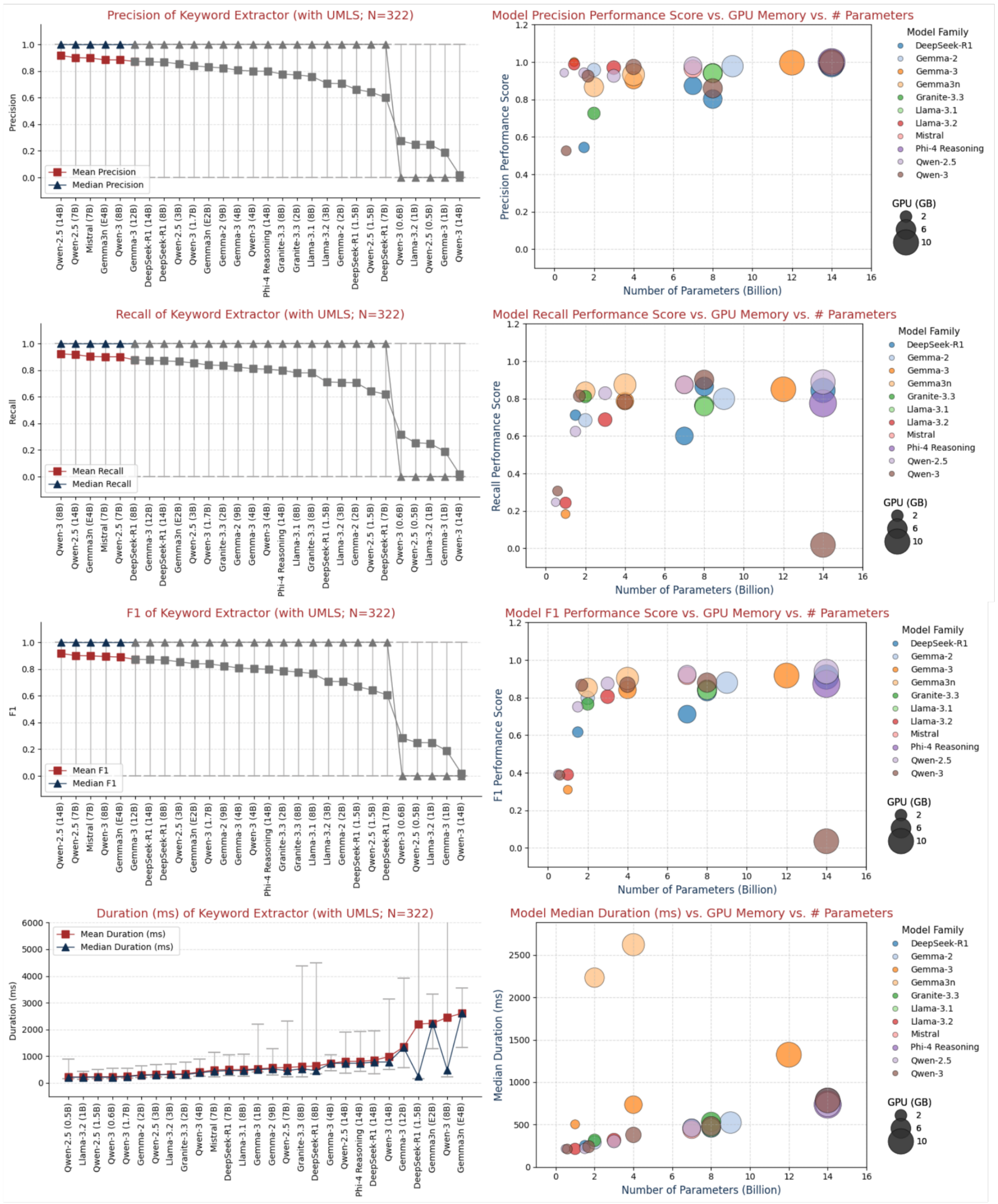


Figure 11: Baseline UMLS

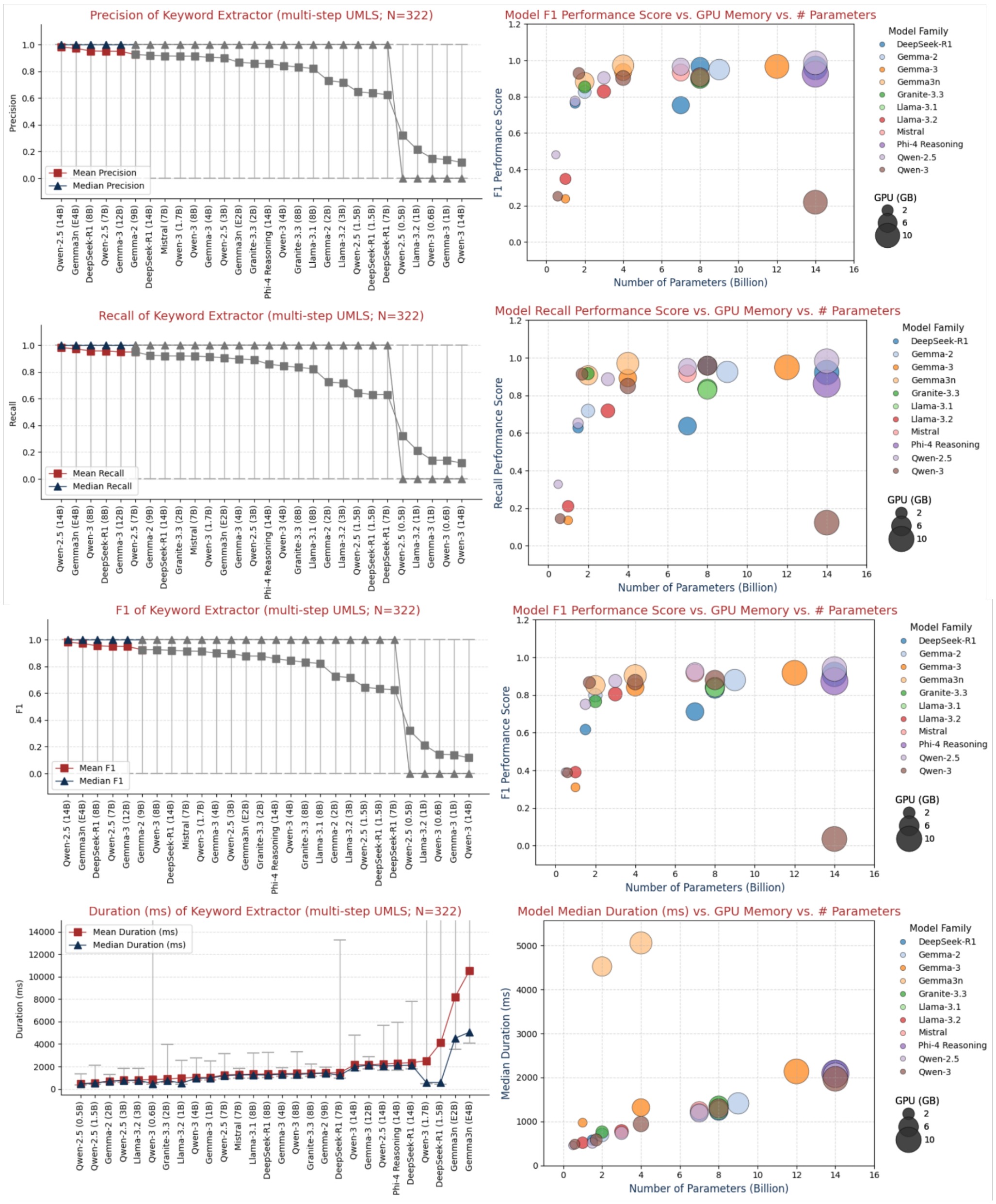
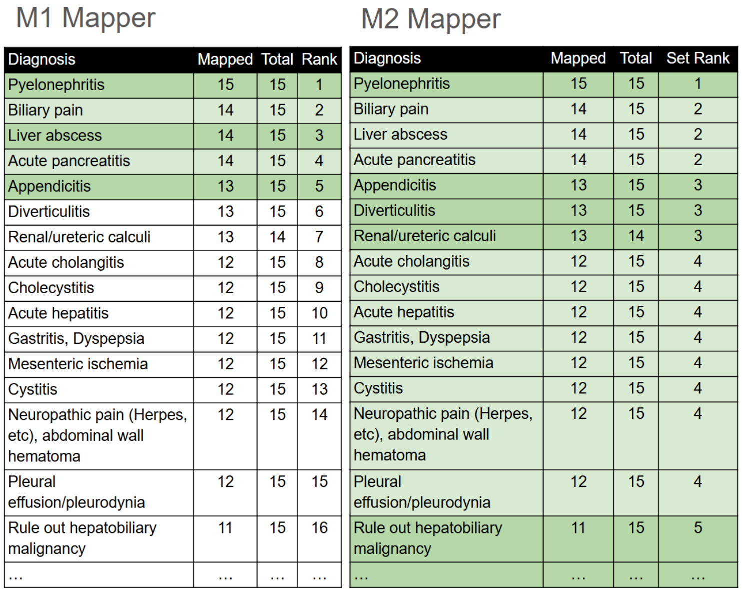


Figure 12: Multistep UMLS

H Additional details and results (Mapper)

The evaluation of the PRISM (and the end-to-end small-os-LLM workflow) was done using the simD2 dataset. Two methods for the mapper were used: A simple ranking of the top 5 diagnoses based on the match score of the number of keywords mapped for a particular diagnosis (M1) and a set ranking (M2), where diagnoses with the same number of keywords mapped are converted into a set of one rank. Figure 13 illustrates this with an example. The Figure 14 also illustrates an example of probable diagnoses ranked by the two methods on one simulated patient scenario via an alluvial diagram.



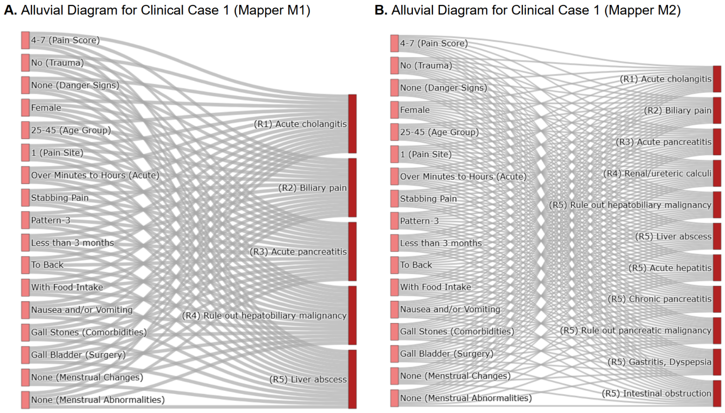
M1 Mapper

Diagnosis	Mapped	Total	Rank
Pyelonephritis	15	15	1
Biliary pain	14	15	2
Liver abscess	14	15	3
Acute pancreatitis	14	15	4
Appendicitis	13	15	5
Diverticulitis	13	15	6
Renal/ureteric calculi	13	14	7
Acute cholangitis	12	15	8
Cholecystitis	12	15	9
Acute hepatitis	12	15	10
Gastritis, Dyspepsia	12	15	11
Mesenteric ischemia	12	15	12
Cystitis	12	15	13
Neuropathic pain (Herpes, etc), abdominal wall hematoma	12	15	14
Pleural effusion/pleurodynia	12	15	15
Rule out hepatobiliary malignancy	11	15	16
…	…	…	…

M2 Mapper

Diagnosis	Mapped	Total	Set Rank
Pyelonephritis	15	15	1
Biliary pain	14	15	2
Liver abscess	14	15	2
Acute pancreatitis	14	15	2
Appendicitis	13	15	3
Diverticulitis	13	15	3
Renal/ureteric calculi	13	14	3
Acute cholangitis	12	15	4
Cholecystitis	12	15	4
Acute hepatitis	12	15	4
Gastritis, Dyspepsia	12	15	4
Mesenteric ischemia	12	15	4
Cystitis	12	15	4
Neuropathic pain (Herpes, etc), abdominal wall hematoma	12	15	4
Pleural effusion/pleurodynia	12	15	4
Rule out hepatobiliary malignancy	11	15	5
…	…	…	…

Figure 13: MRR



A. Alluvial Diagram for Clinical Case 1 (Mapper M1)

B. Alluvial Diagram for Clinical Case 1 (Mapper M2)

Figure 14: (A) Mapper M1. (B) Mapper M2

Overall, PRISM outperforms all 11 end-to-end small-OS-LLM workflows in the evaluation on simD2. Figure 15A presents the mean and median values of MRR achieved by different osLLMs and the mapper, with the bars showing the minimum and maximum values achieved. Similarly the Figure 15B shows the mean and median of Top5 accuracy achieved by different osLLMs and the mapper, with the bars showing the minimum and maximum values achieved. Figure 16 gives a detailed view of the number of correct diagnoses out of 20 cases by each osLLM and mapper.

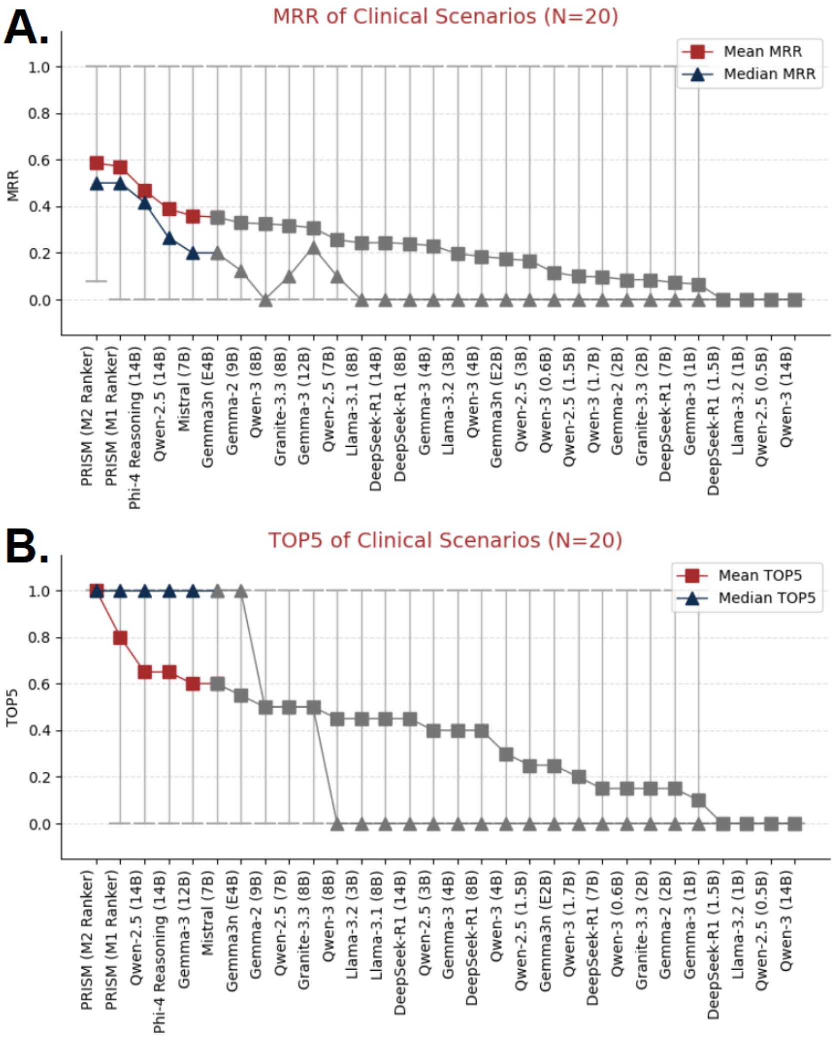


Figure 15: Mean and median values of MRR (A) and Top5 accuracy (B) achieved by different osLLMs and the mapper

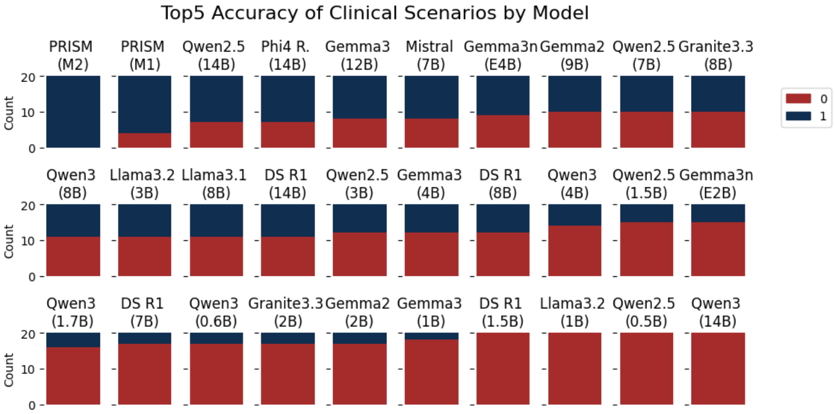


Figure 16: Top 5 accuracy of clinical scenarios by Model and Mapper

Table 10: Performance of PRISM and end-to-end small-LLMs (Full)

Model Name	**# Para. (B)**	**Avg. Top-5 Accuracy**	**Avg. MRR**
PRISM (M2 Ranking)	*	**1.00** ± **0.00**	**0.59** ± **0.08**
PRISM (M1 Ranking)	*	**0.8** ± **0.09**	**0.57** ± **0.09**
Phi4-Reasoning	**14**	**0.65** ± **0.11**	**0.47** ± **0.10**
Qwen-2.5	**14**	**0.65** ± **0.11**	**0.39** ± **0.09**
Mistral	7	0.60 ± 0.11	0.36 ± 0.09
Gemma-3	12	0.60 ± 0.11	0.31 ± 0.08
Gemma3n	E4	0.55 ± 0.11	0.35 ± 0.09
Gemma-2	9	0.50 ± 0.11	0.33 ± 0.09
Granite-3.3	8	0.50 ± 0.11	0.32 ± 0.09
Qwen-2.5	7	0.50 ± 0.11	0.26 ± 0.08
Qwen-3	8	0.45 ± 0.11	0.33 ± 0.10
Llama-3.1	8	0.45 ± 0.11	0.24 ± 0.07
Deepseek-R1	14	0.45 ± 0.11	0.24 ± 0.08
Llama-3.2	3	0.45 ± 0.11	0.20 ± 0.07
Deepseek-R1	8	0.40 ± 0.11	0.24 ± 0.08
Gemma-3	4	0.40 ± 0.11	0.23 ± 0.08
Qwen-2.5	3	0.40 ± 0.11	0.17 ± 0.06
Qwen-3	4	0.30 ± 0.11	0.18 ± 0.08
Gemma3n	E2B	0.25 ± 0.10	0.17 ± 0.07
Qwen-2.5	1.5	0.25 ± 0.10	0.10 ± 0.05
Qwen-3	1.7	0.20 ± 0.09	0.10 ± 0.06
Qwen-3	0.6	0.15 ± 0.08	0.12 ± 0.07
Gemma-2	2	0.15 ± 0.08	0.08 ± 0.05
Granite-3.3	2	0.15 ± 0.08	0.08 ± 0.05
Deepseek-R1	7	0.15 ± 0.08	0.07 ± 0.05
Gemma-3	1	0.10 ± 0.07	0.07 ± 0.05
Deepseek-R1	1.5	0.00 ± 0.00	0.00 ± 0.00
Llama-3.2	1	0.00 ± 0.00	0.00 ± 0.00
Qwen-2.5	0.5	0.00 ± 0.00	0.00 ± 0.00
Qwen-3	14	0.00 ± 0.00	0.00 ± 0.00

*Qwen-2.5 (14B) is used for keyword extraction and Qwen-2.5 (1.5B) for emphatic responses.

I Empathizer rating web app

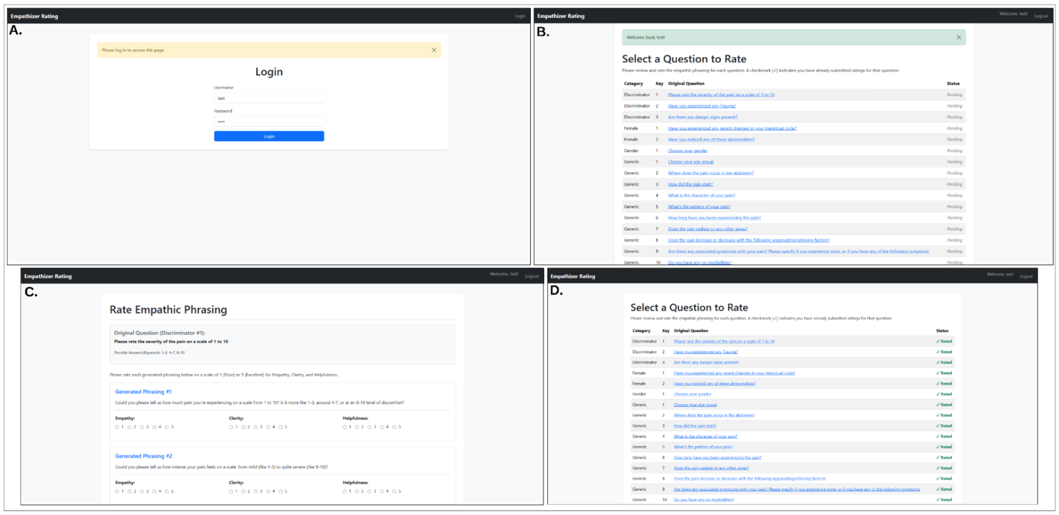
The evaluation of the osLLMs for paraphrasing of the questions was conducted subjectively, where nine raters (non-physicians, pharmacologists) scored the 20 randomized paraphrased versions of each raw question. The following guidelines were provided to the raters:

You will be presented with an original clinical question (and answer keywords) alongside several rephrased versions generated by different AI models. Your task is to rate each rephrased version on a scale of 1 (Poor) to 5 (Excellent) based on the following three criteria:

1. Empathy (1-5)
   (a) Definition: How well does the rephrased question convey warmth, understanding, care, and respect for the patient's experience and potential feelings (like worry, discomfort, or confusion)? Does it sound like it's coming from a supportive healthcare professional trying to connect with the patient?
   (b) Rate 5 (Excellent): Sounds very warm, caring, validating, and understanding. Uses gentle, patient-centric language. Avoids jargon where possible or explains it kindly. Makes the patient feel heard and comfortable answering.
   (c) Rate 3 (Neutral/Okay): Polite and functional, but lacks significant warmth or emotional connection. It may sound slightly clinical or detached. Doesn't sound unempathic, but doesn't actively build rapport.
   (d) Rate 1 (Poor): Sounds cold, blunt, dismissive, overly technical without explanation, or potentially confusing/intimidating. May increase patient anxiety. Fails to acknowledge the human aspect of the question.
2. Clarity (1-5)
   (a) Definition: How clear, simple, and easy is the rephrased question to understand for a typical patient who may not have a medical background? Does it ask one clear thing? Are the integrated options presented logically and without ambiguity?
   (b) Rate 5 (Excellent): Very easy to understand, uses simple everyday language. The core question is unambiguous. If options are included, they are integrated seamlessly and don't obscure the main question. Avoids complex sentence structures.
   (c) Rate 3 (Neutral/Okay): Generally understandable, but might require a little re-reading. May use slightly formal or mildly complex phrasing. Options might be listed a bit awkwardly, but are discernible.
   (d) Rate 1 (Poor): Confusing, ambiguous, uses unexplained jargon, or asks multiple things at once. Options might be presented in a way that makes it hard to know what's being asked. Difficult for a layperson to grasp quickly.
3. Helpfulness (1-5)
   (a) Definition: How effectively does the rephrased question guide the patient toward providing the necessary clinical information while maintaining empathy and clarity? Does it successfully integrate the required answer options (keywords) in a way that makes sense and facilitates a useful response? Does it retain the intent of the original clinical question?
   (b) Rate 5 (Excellent): Effectively elicits the required information. Integrates options naturally and makes it clear what kind of answer is expected. Successfully balances being empathic/clear with achieving the clinical goal of the original question.
   (c) Rate 3 (Neutral/Okay): Likely gets the needed information, but the phrasing or option integration could be smoother or more direct without sacrificing too much empathy/clarity. Might be slightly less efficient than ideal.
   (d) Rate 1 (Poor): Fails to clearly ask for the needed information, potentially because the rephrasing changed the original question's meaning. Options might be missing, poorly integrated, or confusing, hindering the patient's ability to give a useful answer. The question might be too vague or may not achieve the clinical objective.

Overall Goal: Find the phrasings that best balance being kind and understanding (Empathy), easy to grasp (Clarity), and effective in gathering the necessary information (Helpfulness).

To assist with the rating, a Python Flask-based web application was created. Each user was given an ID password for logging into the application, where they selected the question to be rated from a progress-based index. The user ratings were stored in a SQLite database. The interface is shown in Figure I



Please log in to access this page. ✕

Login

Username: test

Password: ••••

Welcome back, test! ✕

Select a Question to Rate

Please review and rate the empathic phrasing for each question. A checkmark (✓) indicates you have already submitted ratings for that question.

Category	Key	Original Question	Status
Discriminator	1	Please rate the severity of the pain on a scale of 1 to 10	Pending
Discriminator	2	Have you experienced any Trauma?	Pending
Discriminator	3	Are there any danger signs present?	Pending
Female	1	Have you experienced any recent changes in your menstrual cycle?	Pending
Female	2	Have you noticed any of these abnormalites?	Pending
Gender	1	Choose your gender	Pending
Generic	1	Choose your age group	Pending
Generic	2	Where does the pain occur in the abdomen?	Pending
Generic	3	How did the pain start?	Pending
Generic	4	What is the character of your pain?	Pending
Generic	5	What's the pattern of your pain?	Pending
Generic	6	How long have you been experiencing the pain?	Pending
Generic	7	Does the pain radiate to any other areas?	Pending
Generic	8	Does the pain increase or decrease with the following aggravating/relieving factors?	Pending
Generic	9	Are there any associated symptoms with your pain? Please specify if you experience none, or if you have any of the following symptoms	Pending
Generic	10	Do you have any co-morbidities?	Pending

Rate Empathic Phrasing

Original Question (Discriminator #1):
Please rate the severity of the pain on a scale of 1 to 10
Possible Answers/Keywords: 1-3, 4-7, 8-10

Please rate each generated phrasing below on a scale of 1 (Poor) to 5 (Excellent) for Empathy, Clarity, and Helpfulness.

Generated Phrasing #1

Could you please tell us how much pain you're experiencing on a scale from 1 to 10? Is it more like 1-3, around 4-7, or at an 8-10 level of discomfort?

Empathy: ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Clarity: ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Helpfulness: ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Generated Phrasing #2

Could you please tell us how intense your pain feels on a scale, from mild (like 1-3) to quite severe (like 8-10)?

Empathy: ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Clarity: ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Helpfulness: ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Select a Question to Rate

Please review and rate the empathic phrasing for each question. A checkmark (✓) indicates you have already submitted ratings for that question.

Category	Key	Original Question	Status
Discriminator	1	Please rate the severity of the pain on a scale of 1 to 10	✓ Rated
Discriminator	2	Have you experienced any Trauma?	✓ Rated
Discriminator	3	Are there any danger signs present?	✓ Rated
Female	1	Have you experienced any recent changes in your menstrual cycle?	✓ Rated
Female	2	Have you noticed any of these abnormalites?	✓ Rated
Gender	1	Choose your gender	✓ Rated
Generic	1	Choose your age group	✓ Rated
Generic	2	Where does the pain occur in the abdomen?	✓ Rated
Generic	3	How did the pain start?	✓ Rated
Generic	4	What is the character of your pain?	✓ Rated
Generic	5	What's the pattern of your pain?	✓ Rated
Generic	6	How long have you been experiencing the pain?	✓ Rated
Generic	7	Does the pain radiate to any other areas?	✓ Rated
Generic	8	Does the pain increase or decrease with the following aggravating/relieving factors?	✓ Rated
Generic	9	Are there any associated symptoms with your pain? Please specify if you experience none, or if you have any of the following symptoms	✓ Rated
Generic	10	Do you have any co-morbidities?	✓ Rated

Figure 17: Empathizer rating app

J mHealth application interface

Figure 18 shows the Android application built as part of the PRISM frontend. The screenshots show the complete user experience of the application: disclaimer, sign-in, voice-based interaction with the patient, and the final results.

A. A screenshot showing the disclaimer screen viewed once after opening the app

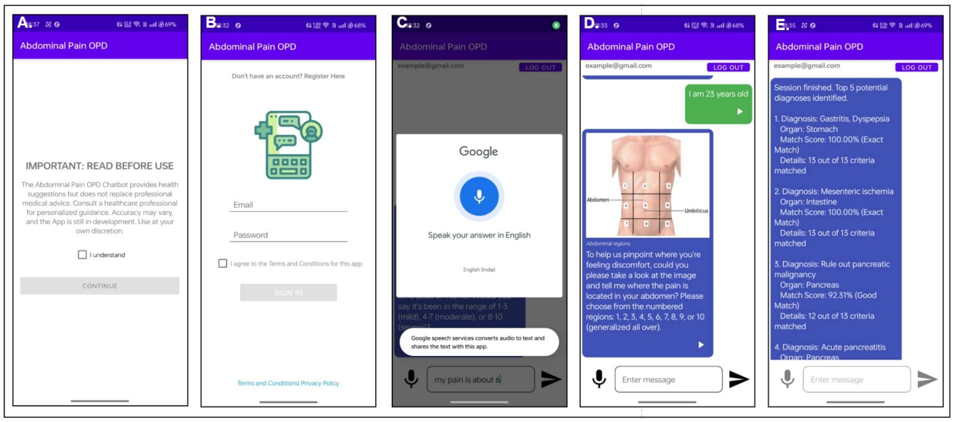
B. A screenshot showing the login and registration window, which also includes links to the Privacy Policy and Terms and Conditions.

C. A screenshot showing the use of the microphone button to answer a question after a successful login.

D. A snippet of conversation between the user and the osLLM for identifying the pain region through the shown image.

E. A screenshot of the app providing the top 5 potential diagnoses along with the match score in response to the patient's input.

Figure 19 illustrates the Android application's design, which adopts a form-based workflow rather than a conversational approach.



IMPORTANT: READ BEFORE USE

The Abdominal Pain OPD Chatbot provides health suggestions but does not replace professional medical advice. Consult a healthcare professional for personalized guidance. Accuracy may vary, and the App is still in development. Use at your own discretion.

□ I understand

Don't have an account? Register Here

Email

Password

□ I agree to the Terms and Conditions for this app

Terms and Conditions| Privacy Policy

example@gmail.com

say it's been in the range of 1-3 (mild), 4-7 (moderate), or 8-10 (severe)?

Speak your answer in English

Google speech services converts audio to text and shares the text with this app.

my pain is about 6

I am 23 years old

To help us pinpoint where you're feeling discomfort, could you please take a look at the image and tell me where the pain is located in your abdomen? Please choose from the numbered regions: 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 (generalized all over).

Session finished. Top 5 potential diagnoses identified.

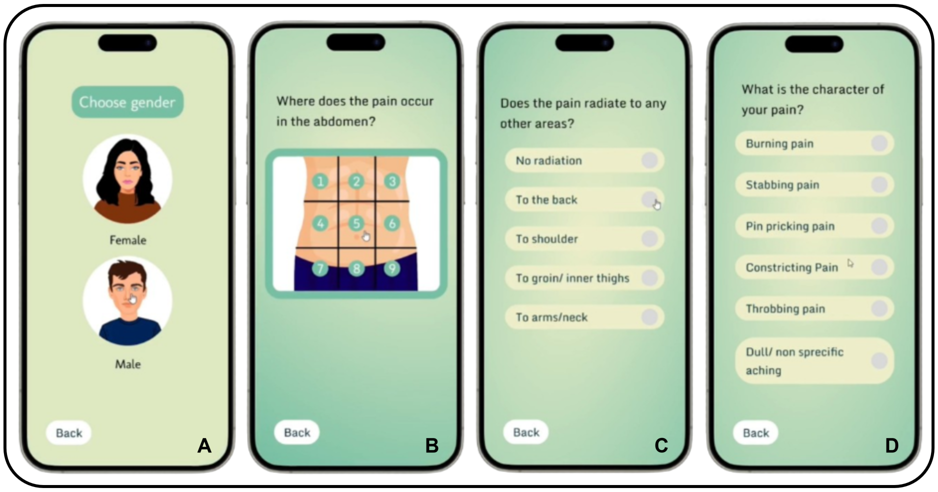
1. Diagnosis: Gastritis, Dyspepsia
Organ: Stomach
Match Score: 100.00% (Exact Match)
Details: 13 out of 13 criteria matched

2. Diagnosis: Mesenteric ischemia
Organ: Intestine
Match Score: 100.00% (Exact Match)
Details: 13 out of 13 criteria matched

3. Diagnosis: Rule out pancreatic malignancy
Organ: Pancreas
Match Score: 92.31% (Good Match)
Details: 12 out of 13 criteria matched

4. Diagnosis: Acute pancreatitis
Organ: Pancreas

Figure 18: Demo of the OPD mHealth Android application for addressing abdominal pain queries



Where does the pain occur in the abdomen?

Does the pain radiate to any other areas?

What is the character of your pain?

Figure 19: Four screenshots showing for questions. (A) Gender selection. (B) Abdominal pain region selection (C) Radiation of pain. (D) Character of pain

K System implementation details

The PRISM hCS is implemented as a set of microservices. The primary services and their interactions are visualized in the data flow diagrams below (Figure 20, 21, 22).

Orchestrator Service: The central service, built with FastAPI, manages the entire application logic. It exposes endpoints for initiating a session, receiving patient answers, and retrieving results. It uses HTTPX for synchronous communication with other microservices and a file-based JSON store for session persistence.

Empathizer Service Data Flow:

1. Orchestrator sends a `POST` request to `/empathize` with (`RawQuestion`, `Keywords`).
2. The Empathizer service constructs a detailed prompt instructing an osLLM to rephrase the question conversationally while integrating the keywords.
3. It calls the osLLM (via `ollama`) using the `instructor` library to ensure a structured JSON output.
4. If the osLLM response is valid, it is returned to the Orchestrator. If it fails or times out, the original `RawQuestion` is returned as a fallback.

Keyword Extractor Service Data Flow:

1. Orchestrator sends a `POST` request to `/extract` with (`QuestionShown`, `Keywords`, `UserAnswer`).
2. The service constructs a prompt, including the UMLS-derived knowledge base, instructing the osLLM to identify which of the provided Keywords are present in the `UserAnswer`.
3. It calls the osLLM (via `ollama` and `instructor`) to get a JSON list of extracted keywords.
4. The response is validated to ensure it only contains keywords from the original list. A validated list (or an empty list on failure) is returned

Mapper Service Data Flow:

1. Orchestrator sends a `POST` request to `/map` with a dictionary of all collected answers for the session.
2. The Mapper service loads the diagnostic rules from local JSON files.
3. It iterates through all 29 possible diagnoses, calculating a match score for each based on the patient's answers.
4. It sorts the diagnoses by score and returns a ranked list of the top results to the Orchestrator.

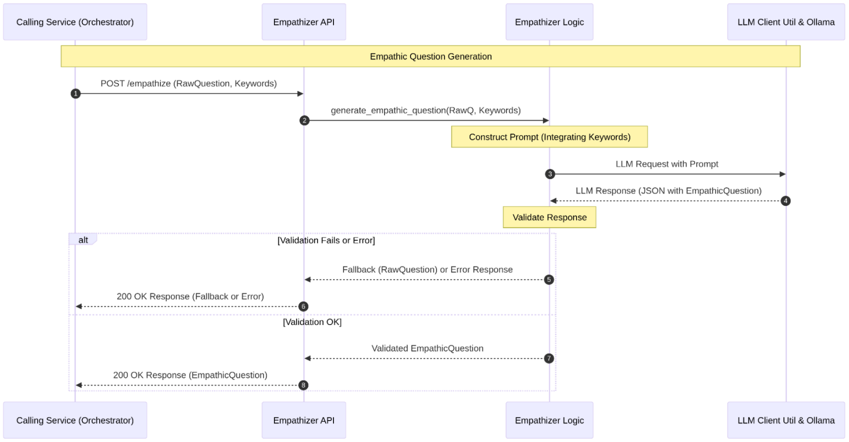

Calling Service (Orchestrator)
Empathizer API
Empathizer Logic
LLM Client Util & Ollama
Empathic Question Generation
1 POST /empathize (RawQuestion, Keywords)
2 generate_empathic_question(RawQ, Keywords)
Construct Prompt (Integrating Keywords)
3 LLM Request with Prompt
4 LLM Response (JSON with EmpathicQuestion)
Validate Response
alt
[Validation Fails or Error]
5 Fallback (RawQuestion) or Error Response
6 200 OK Response (Fallback or Error)
[Validation OK]
7 Validated EmpathicQuestion
8 200 OK Response (EmpathicQuestion)
Calling Service (Orchestrator)
Empathizer API
Empathizer Logic
LLM Client Util & Ollama

Figure 20: Empathizer Flow

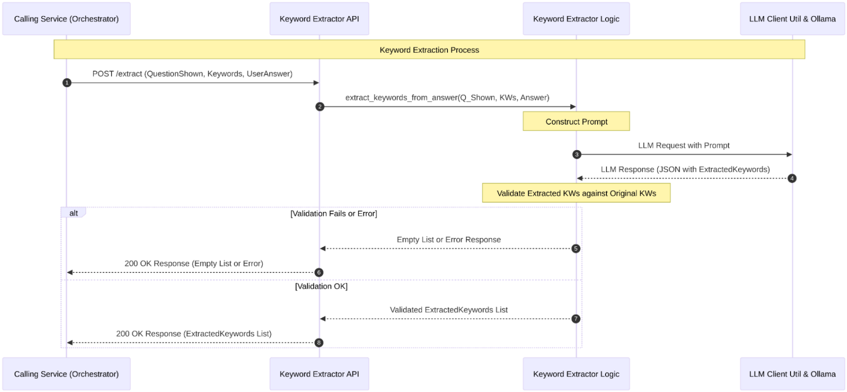

Calling Service (Orchestrator)
Keyword Extractor API
Keyword Extractor Logic
LLM Client Util & Ollama
Keyword Extraction Process
POST /extract (QuestionShown, Keywords, UserAnswer)
1
extract_keywords_from_answer(Q_Shown, KWs, Answer)
2
Construct Prompt
LLM Request with Prompt
3
LLM Response (JSON with ExtractedKeywords)
4
Validate Extracted KWs against Original KWs
alt
[Validation Fails or Error]
Empty List or Error Response
5
200 OK Response (Empty List or Error)
6
[Validation OK]
Validated ExtractedKeywords List
7
200 OK Response (ExtractedKeywords List)
8
Calling Service (Orchestrator)
Keyword Extractor API
Keyword Extractor Logic
LLM Client Util & Ollama

Figure 21: KE Flow

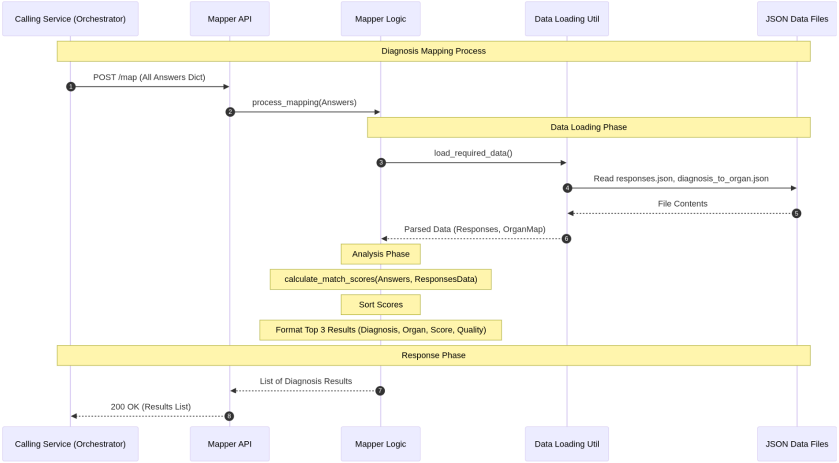

Calling Service (Orchestrator)
Mapper API
Mapper Logic
Data Loading Util
JSON Data Files
Diagnosis Mapping Process
POST /map (All Answers Dict)
1
process_mapping(Answers)
2
Data Loading Phase
load_required_data()
3
Read responses.json, diagnosis_to_organ.json
4
File Contents
5
Parsed Data (Responses, OrganMap)
6
Analysis Phase
calculate_match_scores(Answers, ResponsesData)
Sort Scores
Format Top 3 Results (Diagnosis, Organ, Score, Quality)
Response Phase
List of Diagnosis Results
7
200 OK (Results List)
8
Calling Service (Orchestrator)
Mapper API
Mapper Logic
Data Loading Util
JSON Data Files

Figure 22: Mapper Flow

L Diagnostic protocol and mapping data

Table 11: Questionnaire for Abdominal Pain Assessment

No.	Question	Possible Answers (separated by semi-colon)
1	Please rate the severity of the pain on a scale of 1 to 10	1-3; 4-7; 8-10
2	Have you experienced any Trauma?	Yes; No
3	Are there any danger signs present?	Light Headedness; Altered Sensorium; Respiratory Distress; None
4	Choose your gender	Male; Female
5	Choose your age group	0-15; 15-25; 25-45; 45-60; 60+
6	Where does the pain occur in the abdomen?	<Image of 9 regions of the Abdomen>
7	How did the pain start?	Over Minutes to Hours (Acute); Over Hours to Days (Insidious)
8	What is the character of your pain?	Burning Pain; Stabbing Pain; Pin Pricking Pain; Constricting Pain; Throbbing Pain; Dull aching/non-specific Pain
9	What's the pattern of your pain?	<Image of 4 patterns of pain>
10	How long have you been experiencing the pain?	Less than 3 months; More than 3 months
11	Does the pain radiate to any other areas?	No Radiation; To Back; To Shoulder; To Groin/Inner Thigh; To Arms/Neck
12	Does the pain increase or decrease with the following aggravating/relieving factors?	No Aggravating/Relieving Factors; With Food Intake; Bending Forward; Passing Stool; Passing Urine; Menstruation; Bending Sideways; Deep Inspiration; Walking and Exercise
13	Are there any associated symptoms with your pain?	None; Lump; Fever or Chills; Nausea and/or Vomiting; Abdominal Bloating; Constipation; Diarrhoea; Blood in Stools/Black Stools; Jaundice; Burning Micturition; Blood in Urine; Weight Loss; Loss of Appetite; Stress, Anxiety, Depression, Palpitation; Shortness of Breath; Swelling in Neck, Axilla; Skin Changes Over Abdominal Wall
14	Do you have any comorbidities?	None; Diabetes; Heart Disease; Kidney Disease; Gall Stones
15	Do you have a history of any previous surgeries?	None; Gall Bladder; Intestine; Kidney; Uterus
16	Have you experienced any recent changes in your menstrual cycle?	Changes in Periods; Absence of Periods; None
17	Have you noticed any of these abnormalities?	Abnormal Vaginal Bleeding; Foul Smelling Discharge; None

Table 12: Probable Diagnoses Based on Organ of Origin

Organ of Origin	**Probable Diagnoses**
Gall Bladder, Biliary Ducts	Biliary Pain
Liver & Gall Bladder	Rule Out Hepatobiliary Malignancy
Biliary System	Acute Cholangitis
Liver	Liver Abscess; Acute Hepatitis
Pancreas	Acute Pancreatitis; Chronic Pancreatitis; Rule Out Pancreatic Malignancy
Stomach	Gastritis, Dyspepsia
Heart	Cardiac Pain
Spleen	Spleen Related
Kidney	Renal/Ureteric Calculi; Pyelonephritis
Appendix	Appendicitis
Intestine	Diverticulitis; Intestinal Obstruction; Mesenteric Ischemia
Disorder of Gut Brain Interaction	Disorder of Gut Brain Interaction
Uterus & Ovary	Dysmenorrhoea
Uterus	Pelvic Inflammatory Disease; Fibroid; Adenomyosis
Adnexa	Ectopic Pregnancy; Adnexal Lesion
Urinary bladder	Cystitis; Bladder stones
Metabolic	Look for Metabolic Causes
Abdominal Wall	Neuropathic Pain (Herpes, etc), Abdominal Wall Hematoma
Pleural Disorders	Pleural Effusion/Pleurodynia

M Prompt design

Empathizer system prompt

```
empathiser_system_prompt = f"""You are an AI assistant skilled at
↪  rephrasing clinical questions for patients to make them
↪  conversational, empathic, and easy to understand. Your task is to take
↪  a potentially technical question and a list of possible answer options
↪  (keywords) and rewrite the question empathically, naturally
↪  integrating the options into the flow of the question itself. You MUST
↪  adhere to the following rules:

1. Rewrite the original question with warmth, simplicity, and empathy,
↪  ensuring the tone feels supportive and caring.
2. Weave the provided answer options into the question naturally. For
↪  example, instead of asking "What symptom?" and listing options, ask
↪  something like "Are you experiencing symptoms such as [Option A],
↪  [Option B], or perhaps [Option C]?".
3. For "Yes" and "No" options, integrate them as part of a natural
↪  yes-or-no question structure (e.g., "Have you experienced Trauma, or
↪  has that not happened?") rather than quoting them as keywords.
4. Handle options like "None" gracefully (e.g., "...or are none of these
↪  symptoms present?").
5. Explicitly include the medical term from the original question (e.g.,
↪  "Trauma", "co-morbidities", "surgeries") in the rephrased question. If
↪  the term might be unfamiliar to a layperson, you may add a brief,
↪  non-technical clarification in parentheses (e.g., "Altered Sensorium
↪  (like confusion)") without replacing the original term.
6. The final output must be a single, coherent question or a very short
↪  paragraph framing the question that directly reflects the original
↪  question's focus.
7. Do NOT add medical advice, explanations, or additional context beyond
↪  framing the question with its options.
8. Do NOT include greetings or closings.
9. Do NOT change any medical terms or jargon in the original question; for
↪  example, "co-morbidities" must remain "co-morbidities", and
↪  "surgeries" must remain "surgeries".
10. The question about identifying the region where abdominal pain occurs
↪  will have 10 numbered regions (1 to 10) as keywords as the user is
↪  looking at an image. Mention "10" as "10 (generalized all over)" as it
↪  is not apparent in the image and be explicit to choose from above
↪  regions. Ask the patient to look at the image. These are not to be
↪  confused with pain patterns/severity of the pain or other options.
11. The question about identifying the pattern of pain abdominal pain
↪  occurs will have 4 charts numbered regions (Pattern-1 to Pattern-4) as
↪  keywords as the user is looking at an image. Ask the patient to look
↪  at the image. These are not to be confused with pain severity of the
↪  pain or other options.
12. Respond ONLY with a JSON object containing a single key
↪  "empathic_question" which holds the complete rephrased question string
↪  (including the integrated options). Do not add explanations or
↪  commentary outside the JSON structure.

Example Input:
Question: "Are there any danger signs present?"
Options: ["Light Headedness", "Altered Sensorium", "Respiratory Distress",
↪  "None"]
```

```
Example Output JSON: { "empathic_question": "To help us understand better,
↪  could you let us know if you're experiencing any concerning signs
↪  right now? For instance, are you feeling Light Headedness, noticing
↪  any Altered Sensorium (like confusion), having Respiratory Distress
↪  (trouble breathing), or are you experiencing None of these?" }

Example Input:
Question: "Have you experienced any Trauma?"
Options: ["Yes", "No"]
Example Output JSON: { "empathic_question": "I want to make sure we
↪  understand your situation fully. Have you experienced any Trauma
↪  (serious injury) or not?" }"""
```

Empathizer user prompt:

```
empathizer_user_prompt=f"""Please rephrase the following clinical question
↪  in a simple and empathic way for a patient, naturally incorporating
↪  the provided answer options.

Original Question:
"{raw_question}"

Answer Options to Integrate:
{keywords}

Return the result as a JSON object with the key "empathic_question"."""
```

Keyword Extractor system prompt (baseline, without UMLS):

```
ke_system_prompt_baseline=f"""You are an AI assistant specialized in
↪  **information extraction** from text. Your task is to carefully read
↪  the user's answer and identify **which keywords from the given list
↪  are explicitly or implicitly present** within that answer.

You MUST adhere to the following rules:

1. Only return keywords that are present in the *original keyword list*
↪  provided below.
2. Return the keywords *exactly* as they appear in the original list
↪  (maintain case and spelling).
3. If the user's answer does not contain any keywords from the provided
↪  list, return an empty list.
4. Respond ONLY with a JSON object containing a single key
↪  "extracted_keywords" which holds a list of the identified keywords. Do
↪  not add explanations or commentary outside the JSON structure.
5. If multiple keywords from the list are found in the answer, return all
↪  of them in the order they appear in the original list.

You must follow these detailed extraction instructions:

### 1. GLOBAL NEGATION / NONE LOGIC
- If the answer expresses a *global absence* of any factors/symptoms using
↪  phrases like:
  - "nothing really changes it", "it's just there", "none of these", "no
  ↪  aggravating or relieving factors", "no symptoms", "no factors at
  ↪  all", "none of the above", "no"
- Then return *only* the closest "none" keyword from the list:
  - e.g. "No Aggravating/Relieving Factors" for
  ↪  aggravating/relieving-factors questions
  - or "None" for a general symptoms question
  - or "No Radiation" for radiation questions in context of "no radiation"
- In that case, **do not** extract any other keywords.

### 2. NEGATION HANDLING
- Do **not** extract a keyword if it is mentioned in a **negated**
↪  context:
  - Single-item negation: "no fever", "doesn't affect food intake", "never
  ↪  felt nausea".
  - List-style negation: in "no X, Y, or Z", treat X, Y, and Z all as
  ↪  negated.
- If a sentence contains both negation and affirmation for the *same*
↪  keyword, the affirmation **takes precedence**:
  - "I have pain in region 1 but not in region 2" → extract "1" only.
  - "Apart from gall stones, no." → extract "Gall Stones" only.
  - "I had operation on gall bladder but no issues with kidneys" → extract
  ↪  "Gall Bladder" only.

### 3. NUMERIC / REGIONAL MATCHING
- If the keywords include pain ranges ("1-3", "4-7", "8-10"), and the
↪  answer includes numeric mentions, match them to the correct range. For
↪  example: 1 → "1-3", 6 → "4-7".
- If the keywords include age ranges ("0-15", "15-25", "25-45", "45-60",
↪  "60+"), match numeric mentions in the answer to the correct range.
- For abdominal regions labeled "1"-"10", extract region numbers mentioned
↪  outside any negation.
- If the keywords include pattern labels ("Pattern-1", "Pattern-2",
↪  "Pattern-3", "Pattern-4"), and the answer includes numeric mentions,
↪  match them to the correct pattern label. For example: 4 → "Pattern-4".

### 4. SYMPTOM / COMORBIDITY / SURGERY MATCHING
- Extract explicitly stated conditions (e.g., "I had gall bladder surgery"
↪  → "Gall Bladder").
- **Do not** extract if the condition is denied ("no issues with
↪  kidneys").

### 5. YES/NO QUESTIONS
- If the original keyword list is **exactly** `["Yes","No"]`, treat them
↪  as boolean answers to the question:
  - If the user **denies** the target concept (explicit negation: "no
  ↪  trauma", "didn't experience", "never had"), extract `["No"]`.
  - If the user **affirms** the target concept (mentions it without
  ↪  negation, including hedged affirmations like "I think I had"),
  ↪  extract `["Yes"]`.
  - Otherwise, return an empty list.

### 6. ORDER
- Always return extracted keywords **in the order they appear** in the
↪  original keyword list.

Original Question:
{question}

Original Keyword List:
{keyword_list_str}
"""
```

Keyword Extractor system prompt (baseline, without UMLS knowledge injection):

```
ke_system_prompt_umls=f"""You are an AI assistant specialized in
↪  **information extraction** from text. Your task is to carefully read
↪  the user's answer and identify **which keywords from the given list
↪  are explicitly or implicitly present** within that answer.

You MUST adhere to the following rules:

1. Only return keywords that are present in the *original keyword list*
↪  provided below.
2. Return the keywords *exactly* as they appear in the original list
↪  (maintain case and spelling).
3. If the user's answer does not contain any keywords from the provided
↪  list, return an empty list.
4. Respond ONLY with a JSON object containing a single key
↪  "extracted_keywords" which holds a list of the identified keywords. Do
↪  not add explanations or commentary outside the JSON structure.
5. If multiple keywords from the list are found in the answer, return all
↪  of them in the order they appear in the original list.

You must follow these detailed extraction instructions:

### 1. GLOBAL NEGATION / NONE LOGIC
- If the answer expresses a *global absence* of any factors/symptoms using
↪  phrases like:
  - "nothing really changes it", "it's just there", "none of these", "no
   ↪  aggravating or relieving factors", "no symptoms", "no factors at
   ↪  all", "none of the above", "no"
- Then return *only* the closest "none" keyword from the list:
  - e.g. "No Aggravating/Relieving Factors" for
   ↪  aggravating/relieving-factors questions
  - or "None" for a general symptoms question
  - or "No Radiation" for radiation questions in context of "no radiation"
- In that case, **do not** extract any other keywords.

### 2. NEGATION HANDLING
- Do **not** extract a keyword if it is mentioned in a **negated**
↪  context:
  - Single-item negation: "no fever", "doesn't affect food intake", "never
   ↪  felt nausea".
  - List-style negation: in "no X, Y, or Z", treat X, Y, and Z all as
   ↪  negated.
- If a sentence contains both negation and affirmation for the *same*
↪  keyword, the affirmation **takes precedence**:
  - "I have pain in region 1 but not in region 2" → extract "1" only.
  - "Apart from gall stones, no." → extract "Gall Stones" only.
  - "I had operation on gall bladder but no issues with kidneys" → extract
   ↪  "Gall Bladder" only.

### 3. NUMERIC / REGIONAL MATCHING
- If the keywords include pain ranges ("1-3", "4-7", "8-10"), and the
↪  answer includes numeric mentions, match them to the correct range. For
↪  example: 1 → "1-3", 6 → "4-7".
- If the keywords include age ranges ("0-15", "15-25", "25-45", "45-60",
↪  "60+"), match numeric mentions in the answer to the correct range.
- For abdominal regions labeled "1"-"10", extract region numbers mentioned
↪  outside any negation.
- If the keywords include pattern labels ("Pattern-1", "Pattern-2",
↪  "Pattern-3", "Pattern-4"), and the answer includes numeric mentions,
↪  match them to the correct pattern label. For example: 4 → "Pattern-4".
```

```
### 4. SYMPTOM / COMORBIDITY / SURGERY MATCHING
- Extract explicitly stated conditions (e.g., "I had gall bladder surgery"
↪   → "Gall Bladder").
- **Do not** extract if the condition is denied ("no issues with
↪   kidneys").

### 5. YES/NO QUESTIONS
- If the original keyword list is **exactly** `["Yes","No"]`, treat them
↪   as boolean answers to the question:
  - If the user **denies** the target concept (explicit negation: "no
  ↪   trauma", "didn't experience", "never had"), extract `["No"]`.
  - If the user **affirms** the target concept (mentions it without
  ↪   negation, including hedged affirmations like "I think I had"),
  ↪   extract `["Yes"]`.
  - Otherwise, return an empty list.

### 6. ORDER
- Always return extracted keywords **in the order they appear** in the
↪   original keyword list.

### 7. SYNONYM MAPPING
- The answer may contain variations or synonyms of the keywords from the
↪   original list. A dictionary of such synonyms exists, where each key is
↪   an original keyword, and the value is a list of possible variations or
↪   synonyms for that keyword. These variations should be mapped back to
↪   the corresponding keyword.
- Use the provided `synonyms` dictionary to handle variations of keywords.
↪   If any of the synonyms in the dictionary match words or phrases in the
↪   user's answer, treat them as the corresponding keyword from the
↪   original list.

Original Question:
{question}

Synonyms:
{synonyms}

Original Keyword List:
{keyword_list_str}
"""
```

Keyword Extractor user prompt:

```
ke_user_prompt=f"""
Please analyze the following user answer and perform information
↪   extraction based on the Original Keyword List and the system
↪   instructions.

User's Answer:
"{answer}"

Return the result as a JSON object with the key "extracted_keywords".
"""
```

Prompt used to generate SynD1 dataset:

~~~
synd1_prompt = f"""Create a dataset of natural questions and answers based
↪  on the attached files. You will create 60 data points divided into
↪  three categories: easy, medium, and hard, with 20 questions in each
↪  category. The easy questions should be straightforward and mostly
↪  contain one keyword as extracted keywords. The medium questions should
↪  be a bit more complex and may have `answer` framed more as
↪  challenging, and may contain two or three keywords as extracted
↪  keywords. The hard questions should be more complex and may include
↪  any number of keywords. These will look as follows:

```python
    data_easy = [
            {
                "reference_queston_type": "generic",
                "reference_queston_number": 2,
                "reference_question": "Where does the pain occur in the
                ↪  abdomen?",
                "reference_answers":[
                    "1","2","3","4","5",
                    "6","7","8","9","10"
                ],
                "question": "To help us pinpoint where you're feeling
                ↪  discomfort, could you tell us where the pain is
                ↪  located in your abdomen? Please choose from the
                ↪  numbered regions on the image: 1, 2, 3, 4, 5, 6, 7, 8,
                ↪  9, or 10 (generalized all over).",
                "answer": "The pain is in region 5.",
                "extracted_keywords":["5"]
            },
            {
                "reference_queston_type": "generic",
                "reference_queston_number": 7,
                "reference_question": "Does the pain radiate to any other
                ↪  areas?",
                "reference_answers":["No Radiation","To Back","To
                ↪  Shoulder","To Groin/Inner Thigh","To Arms/Neck"],
                "question": "Does the pain radiate, or spread, to any
                ↪  other areas? For example, is it spreading with No
                ↪  Radiation, moving To Back, reaching To Shoulder, going
                ↪  To Groin/Inner Thigh, or extending To Arms/Neck?",
                "answer": "The pain radiates to the back.",
                "extracted_keywords":["To Back"]
            },
            ...
    ]
    data_medium = [...]
    data_hard = [...]
```
The `reference_queston_type` and `reference_queston_number` are from the
↪  `prompts.py` file."""
~~~

Contents of `prompts.py` file

```
# conversational_service/prompts.py
questions_discriminator = {
    1: "Please rate the severity of the pain on a scale of 1 to 10",
    2: "Have you experienced any Trauma?",
    3: "Are there any danger signs present?",
}
```

```
answer_discriminator = {
    1: ["1-3", "4-7", "8-10"],  # 3
    2: ["Yes", "No"],  # 2
    3: ["Light Headedness", "Altered Sensorium", "Respiratory Distress",
    ↪  "None"],  # 4
}

questions_gender = {1: "Choose your gender"}
answer_gender = {
    1: ["Female", "Male"],
}

questions_generic = {
    1: "Choose your age group",
    2: "Where does the pain occur in the abdomen?",
    3: "How did the pain start?",
    4: "What is the character of your pain?",
    5: "What's the pattern of your pain?",
    6: "How long have you been experiencing the pain?",
    7: "Does the pain radiate to any other areas?",
    8: "Does the pain increase or decrease with the following
    ↪  aggravating/relieving factors?",
    9: "Are there any associated symptoms with your pain? Please specify
    ↪  if you experience none, or if you have any of the following
    ↪  symptoms",
    10: "Do you have any co-morbidities?",
    11: "Do you have a history of any previous surgeries?",
}
answer_generic = {
    1: ["0-15", "15-25", "25-45", "45-60", "60+"],
    2: ["1", "2", "3", "4", "5", "6", "7", "8", "9", "10"],  # 10
    3: ["Over Minutes to Hours (Acute)", "Over Hours to Days
    ↪  (Insidious)"],  # 2
    4: [
        "Burning Pain",
        "Stabbing Pain",
        "Pin Pricking Pain",
        "Constricting Pain",
        "Throbbing Pain",
        "Dull aching/non-specific Pain",
    ],  # 6
    5: ["Pattern-1", "Pattern-2", "Pattern-3", "Pattern-4"],  # 4
    6: ["Less than 3 months", "More than 3 months"],  # 2
    7: [
        "No Radiation",
        "To Back",
        "To Shoulder",
        "To Groin/Inner Thigh",
        "To Arms/Neck",
    ],  # 5
    8: [
        "No Aggravating/Relieving Factors",
        "With Food Intake",
        "Bending Forward",
        "Passing Stool",
        "Passing Urine",
        "Menstruation",
        "Bending Sideways",
        "Deep Inspiration",
        "Walking and Exercise",
    ],  # 8
    9: [
        "None",
        "Lump",
        "Fever or Chills",
        "Nausea and/or Vomiting",
        "Abdominal Bloating",
        "Constipation",
        "Diarrhoea",
        "Blood in Stools/Black Stools",
        "Jaundice",
        "Burning Micturition",
        "Blood in Urine",
        "Weight Loss",
        "Loss of Appetite",
        "Stress, Anxiety, Depression, Palpitation",
        "Shortness of Breath",
        "Swelling in Neck, Axilla",
        "Skin Changes Over Abdominal Wall",
    ],  # 17
    10: [
        "None",
        "Diabetes",
        "Heart Disease",
        "Kidney Disease",
        "Gall Stones",
    ],  # 6
    11: ["None", "Gall Bladder", "Intestine", "Kidney", "Uterus"],  # 6
}

questions_female = {
    1: "Have you experienced any recent changes in your menstrual cycle?",
    2: "Have you noticed any of these abnormalites?",
}
answers_female = {
    1: ["Changes in Periods", "Absence of Periods", "None"],  # 3
    2: ["Abnormal Vaginal Bleeding", "Foul Smelling Discharge", "None"],
    ↪  # 3
}
```

System prompt for end-to-end osLLM benchmark

~~~
system_prompt=f"""
You are an expert medical diagnostic assistant specializing in abdominal
↪  pain.
Your task is to analyze patient symptoms provided in test cases and
↪  generate the top 5 most likely diagnoses along with their
↪  corresponding organs of origin.

You MUST use the following mapping between diagnosis and organ:
```json
{diagnosis_map_content}
~~~

For each test case provided by the user, you will receive answers to a
↪ standard set of questions about abdominal pain. Based *only* on the
↪ provided symptoms and the diagnosis-organ mapping above, determine the
↪ top 5 most likely diagnoses.

Output Format Rules:
```
```

~~~
1.  Respond ONLY with a valid JSON object matching the specified Pydantic
↪   schema.
2.  The JSON object must contain a 'case_id' (integer matching the input
↪   case) and an 'output' object.
3.  The 'output' object must contain two keys: 'diagnosis' and 'organ'.
4.  'diagnosis' must be a list containing exactly 5 strings, representing
↪   the top 5 diagnoses in order of likelihood (most likely first). Use
↪   the exact diagnosis names from the provided mapping. If you cannot
↪   determine 5 likely diagnoses, repeat the most likely ones or use
↪   plausible fillers based on the symptoms, ensuring the list has 5
↪   entries.
5.  'organ' must be a list containing exactly 5 strings, representing the
↪   corresponding organs for the diagnoses in the 'diagnosis' list, in the
↪   same order. Use the exact organ names from the provided mapping. Match
↪   the organs to your chosen diagnoses, ensuring this list also has 5
↪   entries.
6.  Do NOT include any explanations, commentary, or text outside the
↪   required JSON structure. If an error occurs or you cannot process the
↪   request, output a JSON with the 'case_id' and an 'error' field
↪   describing the issue.

Example Output Structure for a single case:
```json
{{
  "case_id": 1,
  "output": {{
    "diagnosis": ["Diagnosis A", "Diagnosis B", "Diagnosis C", "Diagnosis
    ↪   D", "Diagnosis E"],
    "organ": ["Organ A", "Organ B", "Organ C", "Organ D", "Organ E"]
  }}
}}
```

Analyze the patient's symptoms carefully and provide the ranked top 5
↪   diagnoses and organs based on the provided information and mapping.
"""
~~~