MEDPATHAI: An Interpretable Diagnostic Agent with Multilingual and Voice-Enabled Input Using Chain-of-Diagnosis Framework

Project Overview

Title:
MedPathAI: Enhancing Interpretable Diagnostics with Chain-of-Diagnosis (CoD)

Problem Addressed (The Gap):
Traditional LLM diagnostics lack interpretability , rely heavily on synthetic data (synthetic-to-real gap) , and exclude users due to text-only interfaces and language barriers (modality/accessibility gap).

Core Solution:
MedPathAI is an enhanced diagnostic agent that grounds the original Chain-of-Diagnosis (CoD) framework on real-world clinical dialogue data (MedDialog) and integrates a multilingual voice interface.

Key Innovations:
Real-World Dataset Integration: Extends the CoD pipeline by incorporating real-world clinical dialogues alongside synthetic data..
Demographic-Aware Reasoning: Uses patient age and gender to guide follow-up questions and diagnosis.

Accessibility:
Supports multilingual, voice-enabled input .

Interpretable Output:
Provides a five-step reasoning chain and explicit confidence distributions for every diagnosis.

Result:
Achieved 71.2% diagnostic accuracy on real-world data with improved efficiency

Architecture diagram

1. User Input: Receives Text/Voice in Any Language for accessibility.

2. Preprocessing: Normalizes symptoms via Language Detection and Auto-Translation to English.

3. Retriever: Identifies top-K candidate diseases using Hybrid Retrieval (BM25 + Dense Embeddings).

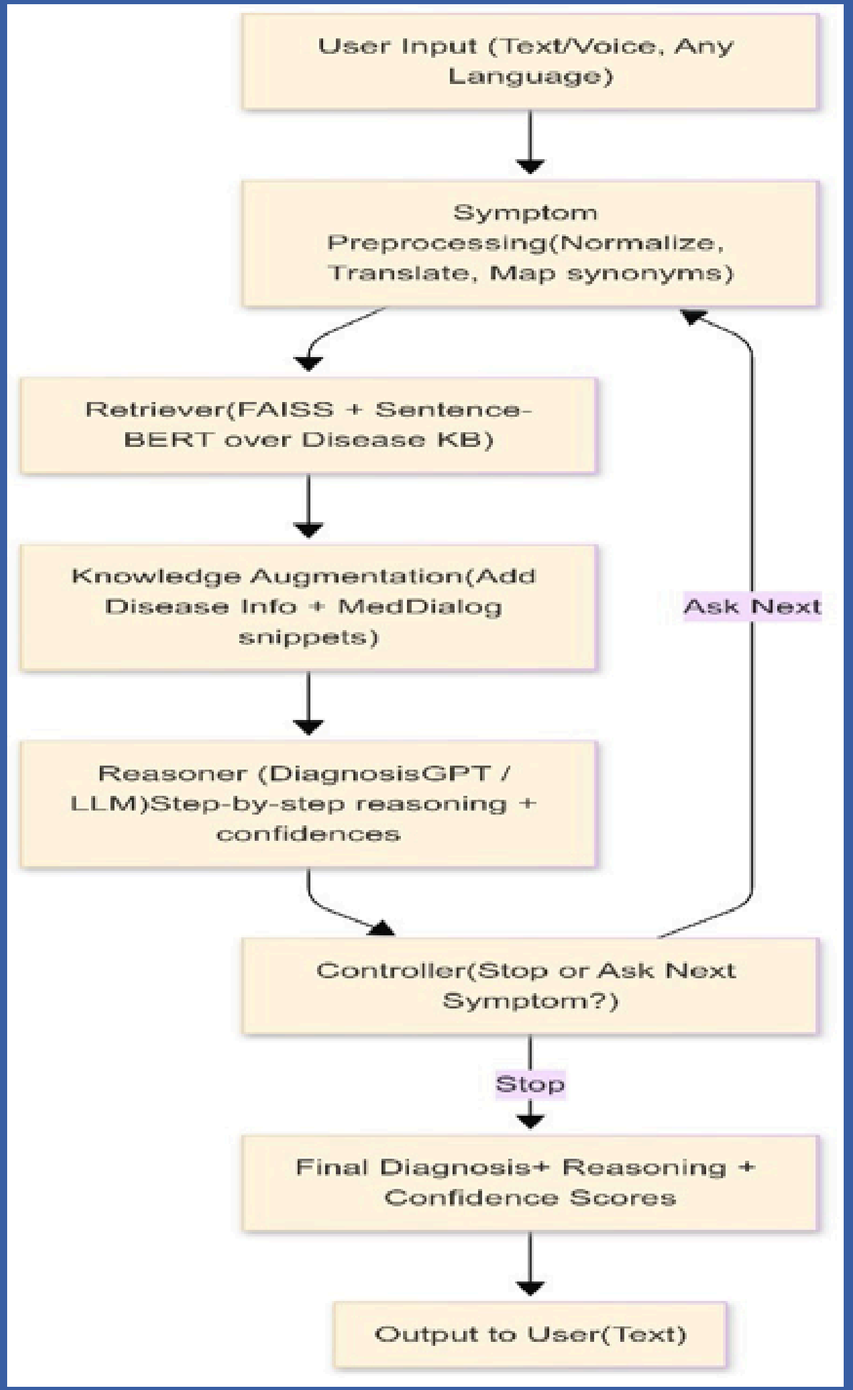
4. Knowledge Augmentation: Grounds reasoning by adding information from the Disease KB and MedDialog (real-world data).

5. Reasoner: Executes CoD steps, performing Diagnostic Reasoning and Confidence Assessment.

6. Controller: The decision block: Diagnoses (Stop) or Inquires (Ask Next). Uses Entropy Reduction to select the most informative next symptom.

7. Final Diagnosis: Compiles the complete, interpretable prediction, reasoning, and confidence scores.

8. Output: Delivers the final report, potentially in the user's native language.



Technology Stack and Tools

Programming Language:

- Python

Large Language Model:

- Qwen2.5-7B-Instruct

Model Fine-Tuning:

- LoRA (PEFT) with Supervised Fine-Tuning using TRL

Retrieval System:

- Sentence-Transformers (all-mpnet-base-v2) for dense embeddings
- BM25 (rank-bm25) for keyword-based retrieval
- FAISS for vector similarity search
- Hybrid retrieval combining BM25 and dense embeddings

Datasets:

- CoD-PatientSymDisease (synthetic diagnostic data)
- MedDialog (real-world clinical conversations)
- Disease_Database (disease–symptom knowledge base)

Reasoning & Logic:

- Confidence scoring
- Entropy-based follow-up question selection

Frontend / UI:

- Streamlit for interactive diagnostic interface
- Voice input support using streamlit_mic_recorder

Deployment:

- Deployed as a web-based Streamlit application

Mathematical Foundation

Core Mathematical Formulas

1. Shannon Entropy (H(c)) : Quantifies the current diagnostic uncertainty or diversity of the confidence distribution C.

$$H(C)=-\sum_{d\in D} C_d \log(c_d)$$

2. Entropy Reduction for Inquiry Selection: The system selects the next symptom () by maximizing the expected reduction in uncertainty (information gain).

$$s_t=\underset{s\in S}{\operatorname{argmax}}(H(C)-E_{Y\sim P(y|s)}[H(C|s,y)])$$

3. Hybrid Retrieval Score: Combines lexical (BM25) and semantic (Dense) matching for efficient candidate recall.

$$score_{hybrid}\,(d,s) = \alpha \,.\, score_{dense}\,(d,s) + (1-\alpha)\,.\, score_{BM25}\,(d,s)$$

Mathematical Foundation

Algorithm Explanations

1. Decision Making (Controller): Determines the next action (Diagnose or Inquire) based on the maximum confidence (C_{max}) and a threshold (τ)

$$A_{next} = \begin{cases} \text{Diagnose}(d_{max}) , & \text{if } C_{max} \geq \tau \text{ and } n_{ing} \geq 3 \\ \text{Inquire}(S_t) , & \text{otherwise} \end{cases}$$

2. Confidence Calculation: The model generates confidence scores, viewed as a softmax distribution over candidate diseases.

$$c_d = \frac{\exp(\text{score}_d / \tau_{\text{temp}})}{\sum_{d' \in D} \exp(\text{score}_{d'} / \tau_{\text{temp}})}$$

LIVE DEMO

Improvement on the paper

Core Improvements

Real-World Validation: First CoD implementation on real clinical dialogues (MedDialog), resolving the synthetic-to-real data gap.
Efficiency: Implemented Entropy Reduction for smart question selection, reducing average inquiries by 29% (to 3.0 rounds).
Accessibility: Integrated multilingual voice interface supporting five languages (English, Hindi, Spanish, French, Mandarin).

Performance Summary

Overall Accuracy: Achieved 71.2% (higher than DiagnosisGPT v3.3 at 69.8% and Original CoD at 68.5%).
Real Data Accuracy: 67.1% (demonstrates generalization from synthetic data's 72.4%).
Efficiency: Average inquiries of 3.0 rounds (significantly lower than the CoD baseline of 4.2).

Challenge & Resolution

Challenge: Integrating the structured CoD framework with ambiguous, real-world MedDialog conversations.
Resolution: Used Hybrid Retrieval and Knowledge Augmentation (with real snippets) to ground the LLM's reasoning.

Future Focus

Limitations: Limited Handling of Rare Diseases and Not Clinically Validated.
Extensions: Expand to support low-resource languages, integrate with EHRs, and use Advanced Reinforcement Learning for question selection.

THANK YOU