Weill Cornell Medicine

A Medical Symptom Diagnosis and Recommendation Generation System Based on LoRA Fine-Tuning and RAG Mechanism

Rongrong Wang, Jinying Xing, Jiazhuo Zheng, Shujun Jiang

Outline

1. Introduction
2. Background & Related Work
3. Data Explanation & Preprocessing
4. Material & Methods
5. Results
6. Discussion
7. Conclusion

Introduction

- **Background**

  The AI in healthcare market size was valued at US$7,679.39 million in 2021 and is expected to grow at a compound annual growth rate (CAGR) of 39.05% during 2022-2027 [1]

- **Problem Identification:**

  Traditional AI diagnosis lacks flexibility for complex symptoms.

- **Proposed Solution:**

  Hybrid system, Fine-tuned LLaMA-3-8B + RAG (NHS knowledge).

- **Impact:**

  Bridges AI and real-world clinical needs.

Weill Cornell Medicine

Artificial Intelligence (AI) in Healthcare Market Size, Share, Trends, Analysis and Forecast by Region, Segment, Offering, Technology and End User, 2022-2027. (2022). *PR Newswire*.

Background & Related Work —— Different from other project

The content of the other project:

https://github.com/codewithsomi/Symptom-Based-Disease-Prediction-Chatbot-Using-NLP/blob/main/Symptom_Based_Disease_Prediction_Chatbot.ipynb

Traditional Approaches and Limitations:

- Limited expressiveness: Inability to process free-text symptom descriptions (e.g., "throbbing headache with light sensitivity").
- Static knowledge: Dependence on fixed datasets, which cannot incorporate new medical guidelines without retraining.
- Minimal contextualization: SVM outputs lack explanatory depth (e.g., "You may have migraines" vs. "Your headache symptoms align with migraines; avoid triggers like caffeine and monitor frequency").
- Inability to process natural language inputs, requiring users to manually map symptoms to predefined options (e.g., selecting "headache" from a dropdown instead of describing "throbbing pain behind my eyes").

Background & Related Work —— Key Innovations

The content of the other project:

https://github.com/codewithsomi/Symptom-Based-Disease-Prediction-Chatbot-Using-NLP/blob/main/Symptom_Based_Disease_Prediction_Chatbot.ipynb

Our work addresses these gaps through four key innovations:

- Natural Language Understanding with LLMs
- Generative Diagnostic Explanations
- Dynamic Knowledge Integration via RAG
- Provides detailed, context-aware medical/treatment advice

Background & Related Work —— Key Contributions

The key innovations of this study are:

- **Open-source adaptation**: This work presents the first integration of a LoRA-fine-tuned LLaMA-3-8B model with the NHS knowledge base, establishing a fully reproducible framework for medical decision support.
- **Dynamic retrieval**: The implementation of FAISS indexing enables millisecond-level knowledge retrieval, ensuring diagnostic recommendations remain synchronized with the most current clinical guidelines.
- **Structured evaluation**: A novel triad evaluation set (41 symptom-diagnosis-recommendation cases) was systematically designed to quantitatively assess both generation quality and factual error rates through standardized metrics.

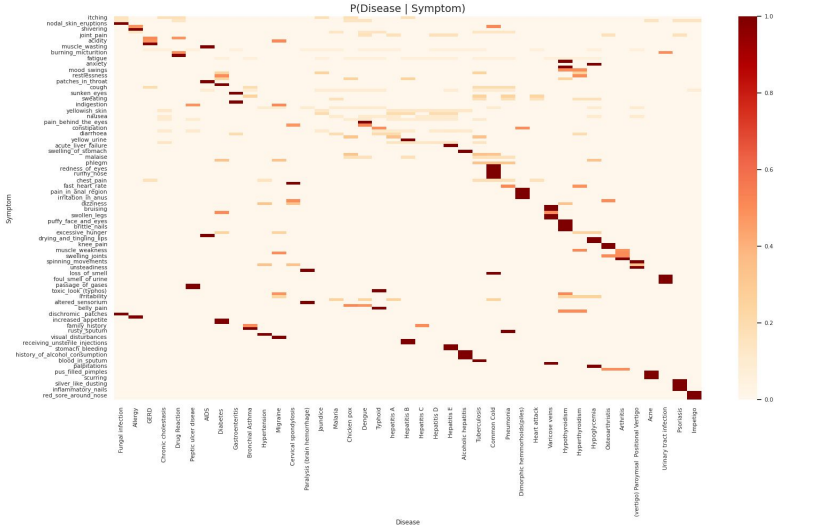
Data Explanation & Preprocessing

Data Source:

Same dataset with the *Symptom Based Disease Prediction Chatbot Using NLP* project.

Contains 4920 combinations of diseases and symptoms with equal number of occurrences of each disease.

Formatting data from *Symcat* increased the number of test entries from 41 to 984



Data Explanation & Preprocessing

Medical Instruction Data for LoRA Tuning:

The training dataset is transformed into structured instruction-output pairs based on symptom, diagnosis, and treatment advice to support LoRA fine-tuning.

Symptom & Disease Data Format

itching	shivering	……	patches	chills	Diagnosis
√		……	√		Fungal infection

Disease & Treatment Advice Data Format

Diagnosis	Treatment Advice
Fungal infection	bath twice
Fungal infection	……
Fungal infection	use clean cloths

Example:
Instruction: "The patient is experiencing itching, skin rash, nodal skin eruptions, dischromic patches. What is the most likely diagnosis and what do you recommend?"
output: "Diagnosis: Fungal infection Advice: - bath twice - use dettol or neem in bathing water - keep infected area dry - use clean cloths"

Advantages: end-to-end prompt style was used for training and inference, in which task description and input content were fused into a natural language instruction. This method is closer to human dialogue expression, which helps to improve the understanding and generation ability of the model in real medical question answering scenarios.

Data Explanation & Preprocessing

Medical Knowledge Base for RAG Enhancement:

- NHS Health A to Z Website

  Collected from the NHS Health A–Z via API and web scraping, covering 60 common diseases.

  Each entry includes structured fields like disease, symptoms, treatment advice.

- OpenEvidence

  For disease entries not fully covered by the NHS website, we manually supplemented the dataset using results retrieved from OpenEvidence's AI-powered medical chatbot.

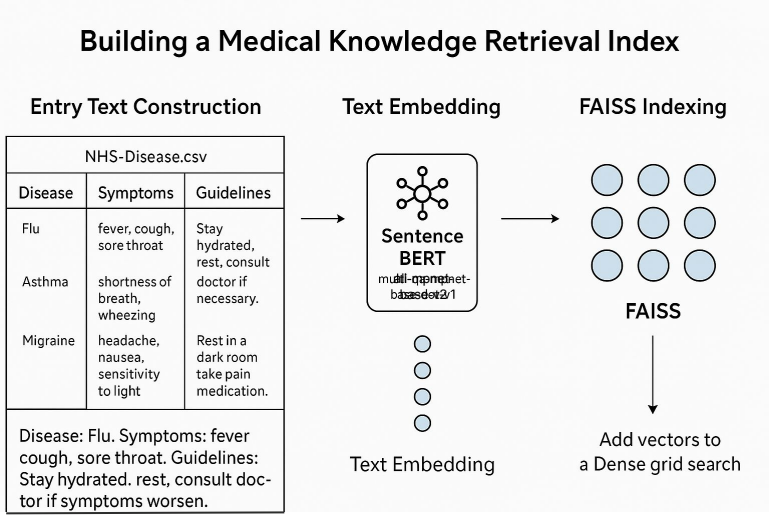
{

id: 17 (disease id in knowledge dataset),
text: “

Disease: hepatitis e.
Symptoms: When symptoms do appear, they can include fever, fatigue, loss of appetite, nausea, vomiting, abdominal pain, jaundice (yellowing of the skin and eyes), dark urine, and pale stools. These symptoms usually appear 2 to 8 weeks after exposure to the virus and can last for several weeks.
Guidelines: The main approach is supportive care, which includes rest, staying hydrated, and eating a healthy diet.

”

}



Data Explanation & Preprocessing

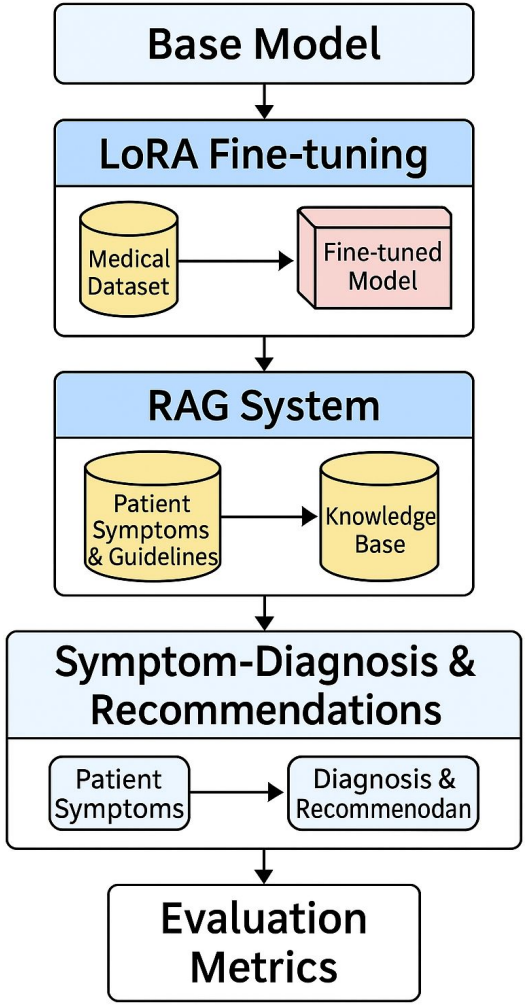
A slight reference to the processing of the generated text before it is subjected to machine evaluation:

- NLTK (Natural Language Toolkit) for word segmentation, stop word processing and basic text cleaning
  - Lowercase: unified uppercase and lowercase to reduce the impact of case difference;
  - Remove punctuation: avoid unnecessary differences caused by symbols;
  - Remove redundant Spaces, line breaks, and unify text formatting;
  - Remove special characters or labels or sentences where the model is not fully completed (truncation due to output limitations).
- Regular expressions extract key information (diagnosed diseases) from formatted text to determine if it is correct.

Material & Methods —— Overall Structure

This project proposes a system designed to predict diseases based on symptom descriptions and to deliver clinically relevant recommendations.

The system comprises two principal components: an open-source large language model fine-tuned using **Low-Rank Adaptation (LoRA)**, and a **Retrieval-Augmented Generation (RAG) module** that leverages external clinical knowledge sources.



Weill Cornell Medicine

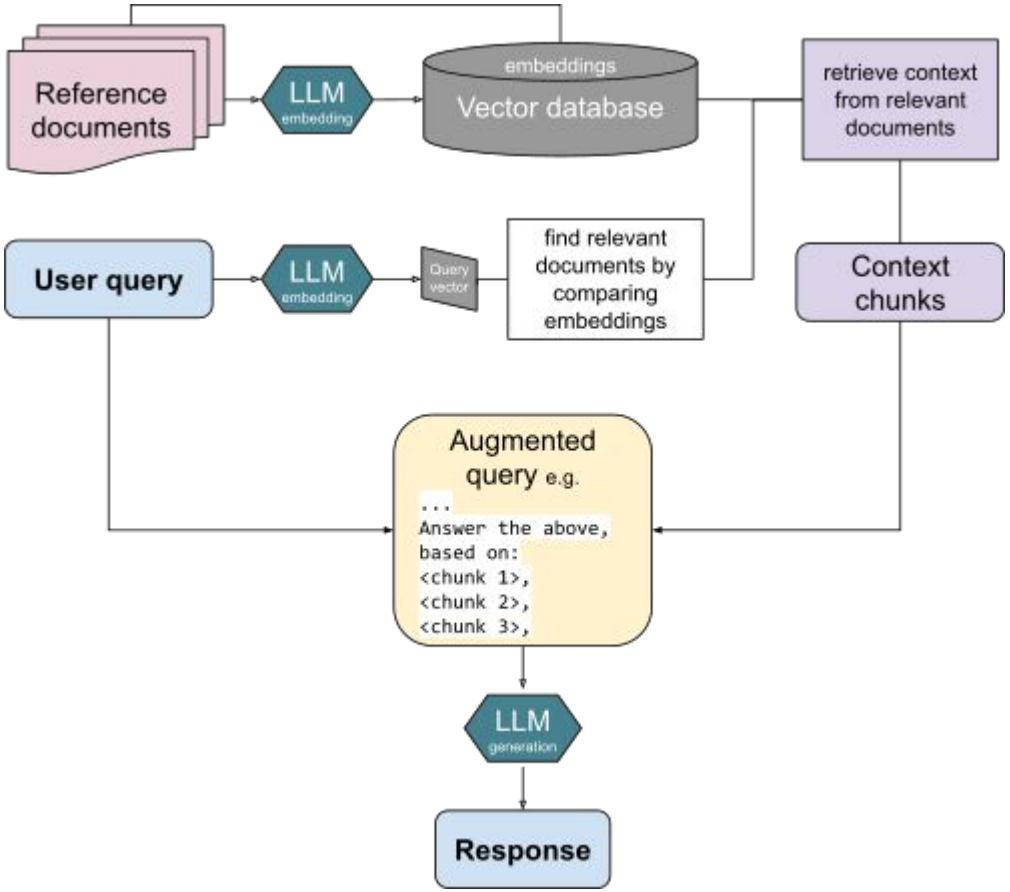
Material & Methods —— Baselines and Enhancement Strategies

- **Baseline Model :**

  Meta - LLaMA-3-8B-Instruct

- **Fine-Tuning Strategy:** LoRA
- **Knowledge Integration:**

  Retrieval-Augmented Generation



Weill Cornell Medicine

Material & Methods —— Experimental Settings

- **LoRA fine-tuning and training parameters**

Parameter-efficient fine-tuning performed on the open-source LLaMA model released by Meta via the Hugging Face platform.

- **LoRA Settings**: rank = 8, alpha = 32, dropout = 0.05
- **Dataset**: Custom-built medical instruction dataset in Alpaca format (instruction / input / output)
- **Training Objective**: Generate diagnostic results and professional advice based on symptom descriptions
- **Training Configuration**:

  Optimizer: AdamW, learning rate = 2e-4

  Epochs: 3; gradient accumulation steps = 4

  Mixed-precision training (fp16) enabled; device_map="auto"

Material & Methods —— Experimental Settings

- **RAG embedding representation and search parameterization**

A medical knowledge base was first constructed, consisting of **60 structured disease entries** derived from official **NHS Conditions** content and manually supplemented data. Each entry contains three components: **disease name**, **typical symptoms**, and **clinical guidelines**, and is formatted in natural language.

For example:

Disease: *Chronic cholestasis*

Symptoms: *Common symptoms include Itching, fatigue, jaundice, dark urine, pale stools, abdominal pain, weight loss, vitamin (A, D, E, K) deficiencies.*

Guidelines: *Treat with UDCA; add obeticholic acid if needed. Use cholestyramine for itching, or rifampin/naltrexone/sertraline if refractory. Ensure vitamin supplementation; consider liver transplant in severe cases.*

Material & Methods —— Experimental Settings

- **RAG embedding representation and search parameterization**

These entries were embedded using the **all-mpnet-base-v2** model from the SentenceTransformers framework . A **FAISS** index was built for dense vector retrieval based on L2 similarity.

During inference, the user's symptom description is encoded into a vector and used to retrieve the top-3 most relevant knowledge entries. Retrieved content is injected into a structured prompt, which is passed to the fine-tuned LLaMA model for response generation.

Generation settings:

- Temperature = 0.7
- Max new tokens = 200
- Nucleus sampling with top_p = 0.9

Material & Methods —— Evaluation Metrics

- **Quantitative evaluation**

  METEOR: Captures synonym-level semantic similarity

  BLEU: Measures n-gram overlap with reference answers

  ROUGE-1 / ROUGE-L: Assesses recall of important content units

  Fuzzy Matching Accuracy: Evaluates approximate string match

Results

Model	METEOR	BLEU	ROUGE (1/L)	Fuzzy Matching Accuracy
Llama3(8b)	0.12	0.17	0.19/0.19	29.88%
Llama3(8b) – LoRA	0.64	0.89	0.90/0.90	91.36%
Llama3(8b) – LoRA – RAG	0.65	0.91	0.92/0.92	95.83%

- **Baseline: LLaMA3 (8B)**

  Poor performance

  severe hallucinations

  cannot handle medical tasks

- **LoRA Fine-tuned**

  Clear, fluent outputs with accurate symptom-diagnosis mapping

- **LoRA + RAG Enhanced**

  More factual and structured answers with knowledge citations

Conclusion - System Summary & Key Results

- Built a **hybrid medical QA system** using **LoRA fine-tuning + RAG retrieval**
- Based on **LLaMA3-8B-Instruct**, designed for **symptom → diagnosis + advice**
- **LoRA**: lightweight fine-tuning, works on a single GPU

  RAG: adds trusted medical knowledge to improve accuracy and explainability
- **Results**:

  Fuzzy Accuracy: 29.88% → 91.36% → **95.83%**

  BLEU, ROUGE, METEOR all significantly improved
- The system is **open-source, low-cost, reproducible**, and easy to deploy in clinical or research settings

Conclusion

Current Limitations

- **Knowledge base is small** (60 NHS entries only)
- **Limited generalizability** in rare or ambiguous symptoms
- **No real-world clinical validation yet**

Future Directions

- Expand the knowledge base to cover **more diseases & treatment cases**
- Introduce **multi-hop reasoning** or causal chain generation
- Explore **integration with real EHR data** and **doctor-in-the-loop evaluation**

Thank you

Weill Cornell Medicine

Weill Cornell Medicine