K S SCHOOL OF ENGINEERING & MANAGEMENT
BENGALURU-560109

(Affiliated to VTU, Belagavi & Approved by AICTE, New Delhi, Accredited by NAAC)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

PROJECT WORK (21ADP76)

TITLE: AI Doctor – A Multi-Modal Medical Diagnostic System

Project Externals
Group No:5
Date: 29/05/2025

NAME	USN
Mohammed Faisal	1KG21AD029
Pranit Chatterjee	1KG21AD037
Praveen A	1KG21AD041
Shriddhi Sinha	1KG21AD045

UNDER THE GUIDANCE

Dr. Pavana H
Assistant Professor,
Dept. of AI&DS, K.S.S.E.M, Bengaluru

CONTENT

- Introduction
- Problem Statement
- Objectives
- Methodology of objective1,2,3
- Results & Discussion
- Conclusion

INTRODUCTION

The AI Doctor project is an innovative initiative designed to improve the primary diagnosis using deep learning techniques across multiple modalities. Our goal is to develop a system that understands complex medical queries, transcribes domain-specific spoken inputs, and analyzes diagnostic images to deliver comprehensive, context-aware assessments. The project addresses key challenges in modern healthcare, including the need for accurate diagnostics, improved interpretability of AI reasoning, and efficient resource utilization.

Phase 2 of the AI Doctor project focuses on integrating visual and textual modalities to enable the system to analyze medical images (e.g., X-rays, MRIs, skin) and generate comprehensive diagnostic narratives. Building on the text-based chain-of-thought foundation established in Phase 1, this phase extends the pipeline to incorporate image understanding, delivering richer, context-aware assessments in clinical settings.

PROBLEM STATEMENT

Modern medical diagnostics are challenged by the fragmented nature of data and the opacity of decision-making processes. Current systems often rely on isolated data sources—textual records, spoken patient interactions, or medical images—each analyzed independently, which limits the ability to form a cohesive and comprehensive understanding of a patient's condition. Additionally, while advanced AI models exist, they frequently lack the transparency and detailed reasoning needed to support clinical decisions, making it difficult for healthcare professionals to trust their outputs.

This project addresses these critical gaps by developing an integrated, multi-modal diagnostic system—AI Doctor—that leverages state-of-the-art deep learning techniques across multi-modalities. By fine-tuning a pre-trained language model with a structured chain-of-thought approach, the project aims to deliver holistic, context-aware, and reliable medical assessments. This integrated solution not only enhances diagnostic accuracy but also provides clear, interpretable reasoning, ultimately supporting better patient care and decision-making in clinical settings.

OBJECTIVES

Objective 1: Text Modality:

- Fine-tune a pre-trained language model—specifically, DeepSeek R1 Distill-Llama 8B—using LoRA-based parameter-efficient training on a Medical Chain-of-Thought (CoT) dataset.

Objective 2: Testing Multi-Modality:

- Evaluate, adapt, and fine-tune a multi-modal LLM (Llama-3.2-11B-Vision-Instruct) on radiology and skin imagery using parameter-efficient training, ensuring both vision and language components are optimized for diagnostic tasks.

Objective 3: Implenting Multi-Modality & User Interface:

- Design and enhance a user-friendly interface that ensures intuitive navigation, improved accessibility, and a seamless user experience for interacting with the multimodal medical chatbot.

METHODOLOGY

Objective 1: Approach for Fine-Tuning the Text Modality:

1.Model and Tokenizer Initialization:

- Setup: Loaded the DeepSeek R1 Distill-Llama 8B model using the FastLanguageModel.from_pretrained() method from the unsloth library.
- Configurations:
  - Max Sequence Length: 2048 tokens to ensure extensive context handling.
  - 4-Bit Quantization: Employed to reduce memory usage and improve inference speed without significant accuracy loss.

2. Custom System Prompt Design:

Chain-of-Thought Integration: Developed a structured prompt that instructs the model to generate detailed chain-of-thought reasoning before providing a final answer.

3.Dataset Preparation:

- **Dataset Used:** The Medical O1 Reasoning SFT dataset from Hugging Face, containing medical questions, detailed reasoning, and final answers.
- **Formatting:** Customized data formatting function to align each entry with the training prompt style, ensuring consistency and effective training signal.

METHODOLOGY

Objective 1: Approach for Fine-Tuning the Text Modality:

Fine-Tuning with LoRA:

- **Parameter Efficiency:** Integrated LoRA adapters using the get_peft_model() function, targeting key transformer layers for efficient adaptation.
- **Training:** Utilized the SFTTrainer from the TRL library to supervise fine-tuning, ensuring that the model learns to generate logical, step-by-step reasoning and concise responses.
-

Inference and Testing:

- **Optimization:** Enabled optimized inference via unsloth's techniques, leading to improved speed (approximately 2x faster) and response quality.
- **Validation:** Conducted preliminary tests to verify that the model generates detailed reasoning and synthesizes it into accurate clinical answers.

METHODOLOGY

Objective 2: Testing Multi-Modality:

1.Model Initialization:

•Initialized the LLaMA-3.2-11B-Vision-Instruct model using the unsloth library.
•Enabled 4-bit quantization to significantly reduce GPU memory usage without compromising performance.
•Applied gradient checkpointing to save memory during backpropagation, allowing training on standard hardware.

2. **Dataset Preparation (Radiology & Dermatology)**

Used two datasets:

- **Radiology:** Chest X-rays / CT scans with expert-written diagnostic reports.
- **Dermatology**: Skin lesion images with expert descriptions or diagnoses.

Converted data into chat-style format:

- **Radiology Prompt:** "You are an expert radiographer. Describe the findings in this image."
- **Dermatology Prompt:** "You are a dermatologist. Describe the skin condition shown in this image."

üAssistant response contained the corresponding diagnostic interpretation.
üThis format replicates real clinical Q&A workflows.

METHODOLOGY

Objective 2: Testing Multi-Modality:

3.Parameter-Efficient Fine-Tuning

Used **Low-Rank Adaptation (LoRA)** to fine-tune efficiently:

- Applied to **vision and language layers**, especially **attention** and **MLP modules**.
- Trained **less than 1 billion parameters**, reducing compute load.

4. Supervised Training Configuration

•Fine-tuning handled using **SFTTrainer**:

- **Batch Size**: 2
- **Gradient Accumulation**: 4 (to simulate batch size = 8)
- **Learning Rate**: 2e-4
- **Precision**: fp16 for faster computation

•Each domain (radiology and dermatology) was trained **separately** using **identical configurations**.

◦

METHODOLOGY

Objective 2: Testing Multi-Modality:

5. Inference Optimization & Evaluation

Post-training optimization with **for_inference()**:

- Enabled **efficient text generation** based on image inputs.

Conducted **evaluation on unseen samples**:

- Checked for **clinical correctness**, **use of accurate terminology**, and **alignment with image content**.

6. Outcome

The model performed well in both domains:

- Showed strong **image grounding**, **structured responses**, and **clinical coherence**.
- **Saved model and tokenizer** for integration into the unified **AI Doctor pipeline**.

METHODOLOGY

Objective 3: Implementing the Multimodal Pipeline & User Interface:

We packaged our fine-tuned text (LoRA-Distill-LLaMA), vision (Llama-3.2-11B-Vision-Instruct), and Whisper-transcribed speech outputs into a single llama-4-scout-17b-16e-instruct model on Groq Cloud. Each input channel feeds its latent representation into dedicated heads within this model, and an attention-based fusion layer learns to weight modalities based on confidence scores.

The combined model was benchmarked and optimized on Groq’s inference hardware. We quantized weights to 4 bits and applied mixed-precision arithmetic, achieving sub-second end-to-end latency. Automated load testing ensured that Gradio API endpoints could sustain concurrent user sessions without degradation.

METHODOLOGY

Objective 3: Implementing the Multimodal Pipeline & User Interface:

User Interface

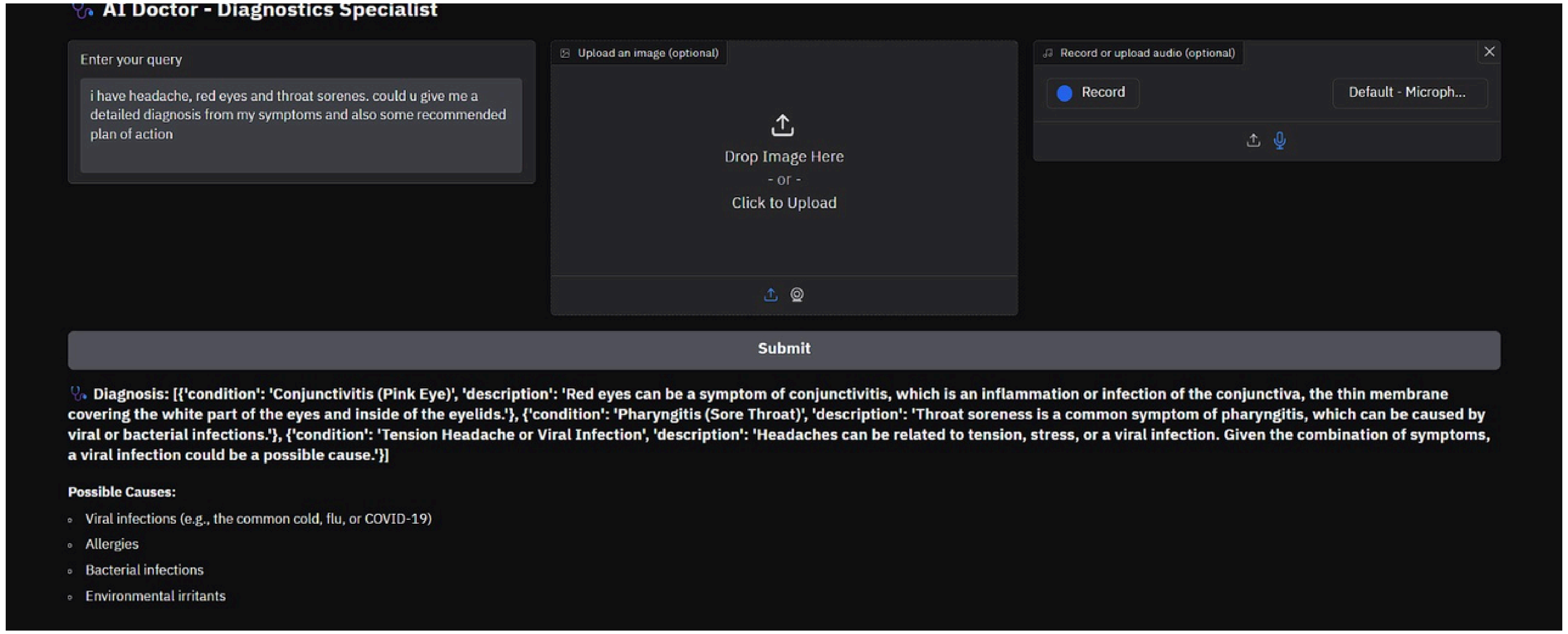
With the multi-modality pipeline stable, we designed a browser-based front end in Gradio. Our interface features three tabs labeled “Text,” “Voice,” and “Image,” each visually distinct but functionally unified:

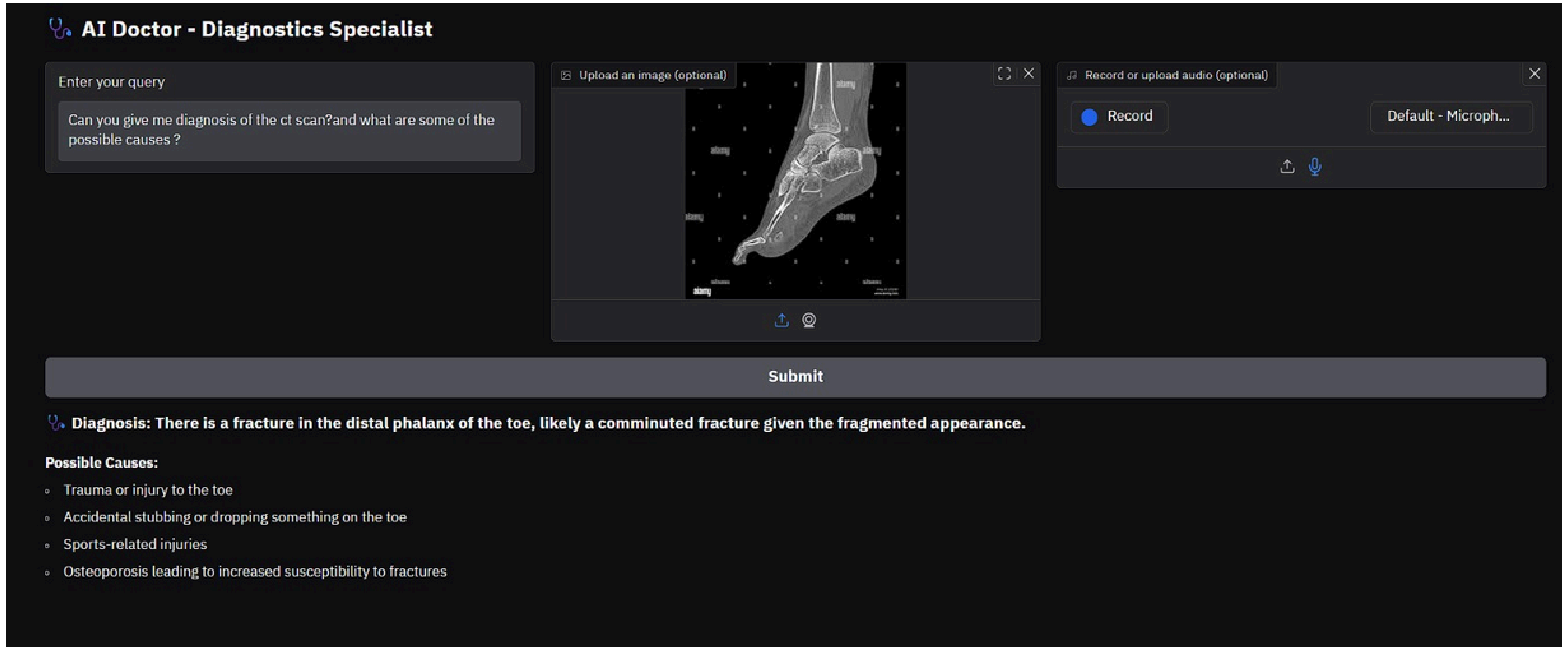
Text Tab: Presents a resizable text area with live character count and optional “Upload Document” function.

Voice Tab: Offers a record/playback widget; pressing “Record” opens a waveform display and live transcription, while “Stop” submits the audio for analysis.

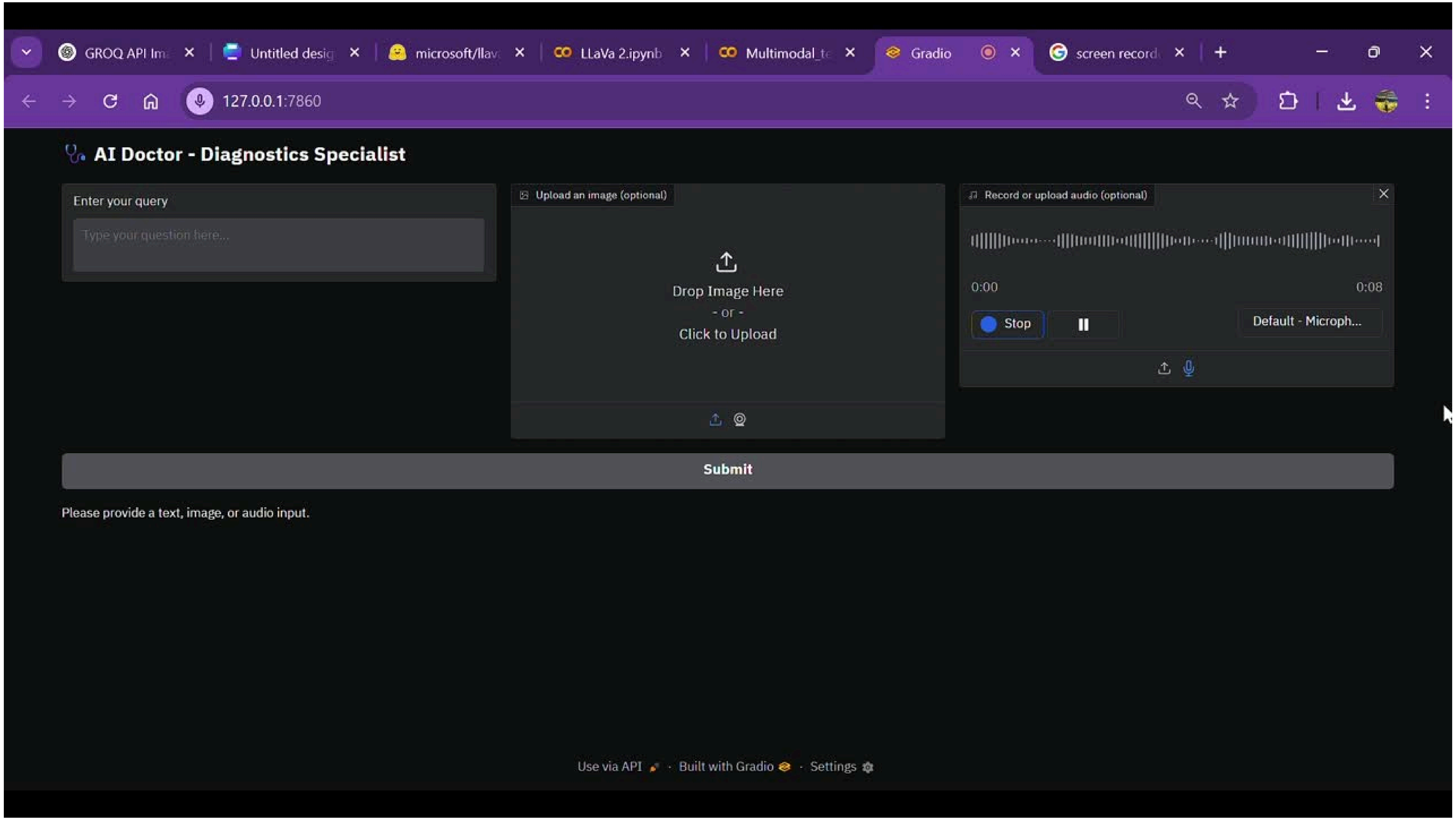
Image Tab: Enables drag-and-drop or click-to-browse image uploads, with an instant thumbnail preview and optional cropping tool.

UI/UX(SCREEN SHOT OF UI)

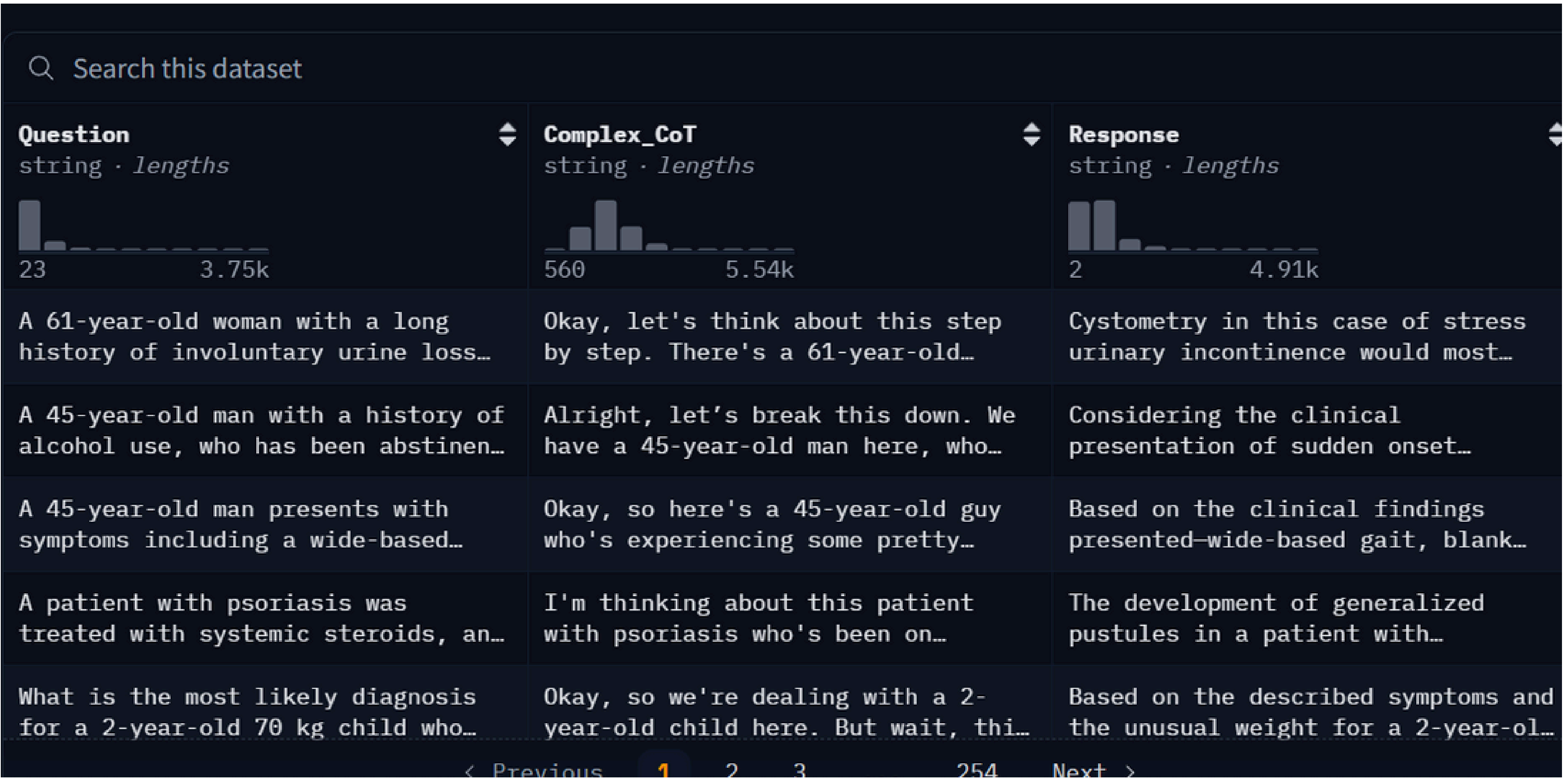




UI/UX(RECORDING OF VOICE MODAL)



DATA SOURCE(LINK / SAMPLE)



Search this dataset

Question string · lengths 23 3.75k	Complex_CoT string · lengths 560 5.54k	Response string · lengths 2 4.91k
A 61-year-old woman with a long history of involuntary urine loss…	Okay, let's think about this step by step. There's a 61-year-old…	Cystometry in this case of stress urinary incontinence would most…
A 45-year-old man with a history of alcohol use, who has been abstinen…	Alright, let's break this down. We have a 45-year-old man here, who…	Considering the clinical presentation of sudden onset…
A 45-year-old man presents with symptoms including a wide-based…	Okay, so here's a 45-year-old guy who's experiencing some pretty…	Based on the clinical findings presented—wide-based gait, blank…
A patient with psoriasis was treated with systemic steroids, an…	I'm thinking about this patient with psoriasis who's been on…	The development of generalized pustules in a patient with…
What is the most likely diagnosis for a 2-year-old 70 kg child who…	Okay, so we're dealing with a 2-year-old child here. But wait, thi…	Based on the described symptoms and the unusual weight for a 2-year-ol…

< Previous 1 2 3 … 254 Next >

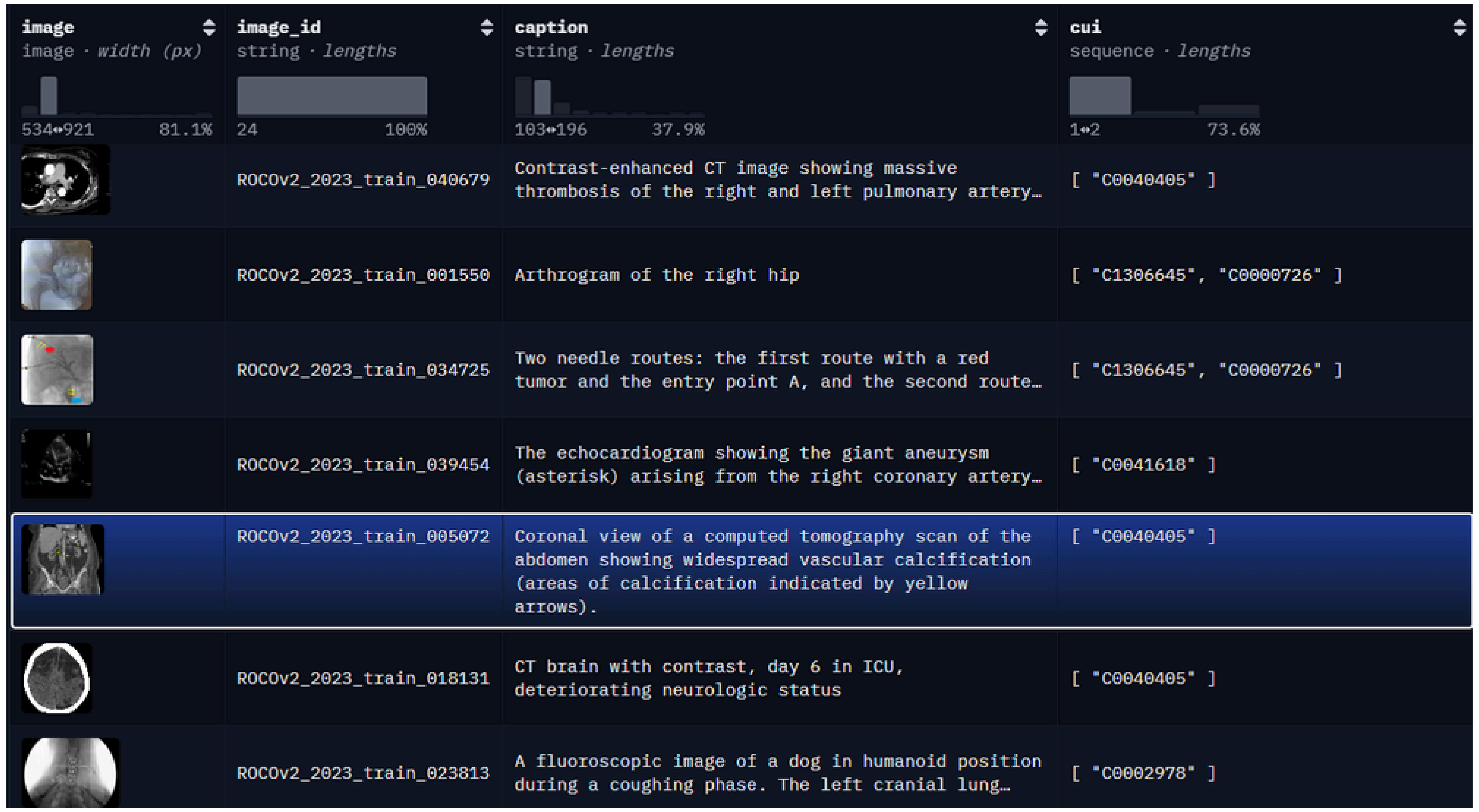
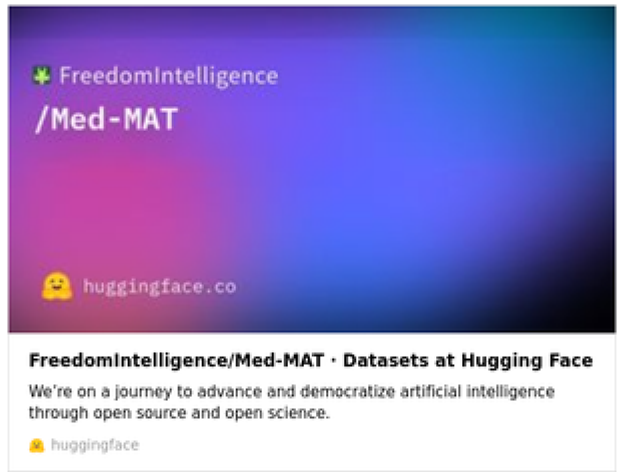
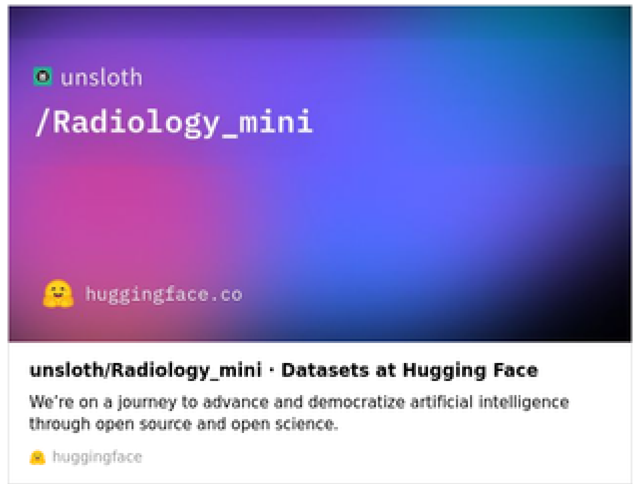


image image · width (px) 534↔921 81.1%	image_id string · lengths 24 100%	caption string · lengths 103↔196 37.9%	cui sequence · lengths 1↔2 73.6%
	ROCOv2_2023_train_040679	Contrast-enhanced CT image showing massive thrombosis of the right and left pulmonary artery…	["C0040405"]
	ROCOv2_2023_train_001550	Arthrogram of the right hip	["C1306645", "C0000726"]
	ROCOv2_2023_train_034725	Two needle routes: the first route with a red tumor and the entry point A, and the second route…	["C1306645", "C0000726"]
	ROCOv2_2023_train_039454	The echocardiogram showing the giant aneurysm (asterisk) arising from the right coronary artery…	["C0041618"]
	ROCOv2_2023_train_005072	Coronal view of a computed tomography scan of the abdomen showing widespread vascular calcification (areas of calcification indicated by yellow arrows).	["C0040405"]
	ROCOv2_2023_train_018131	CT brain with contrast, day 6 in ICU, deteriorating neurologic status	["C0040405"]
	ROCOv2_2023_train_023813	A fluoroscopic image of a dog in humanoid position during a coughing phase. The left cranial lung…	["C0002978"]

Fine-Tuning (Text Part)

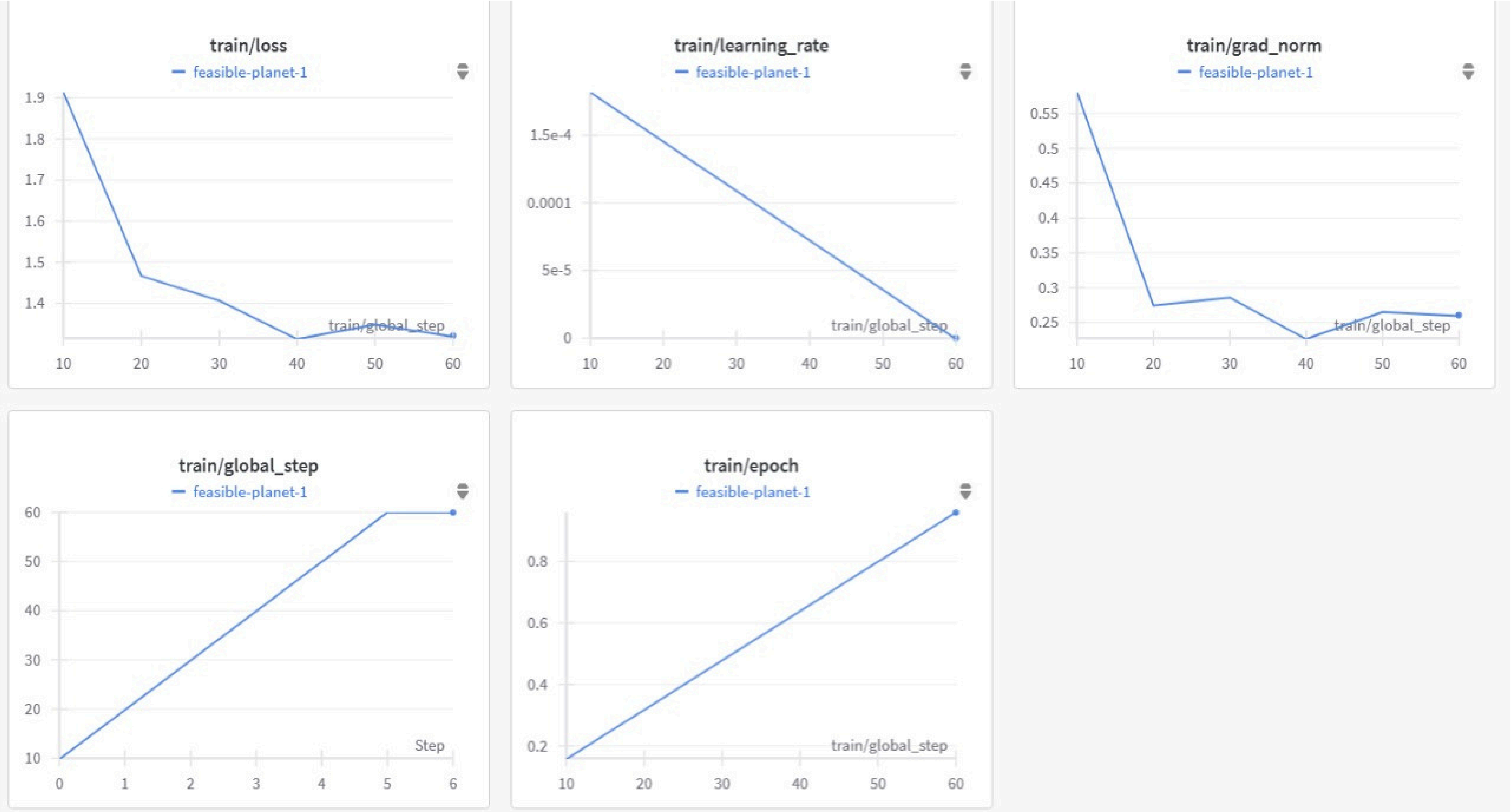


Fine-Tuning (Image Part)



RESULTS(SCREEN SHOT OF OVERALL RESULT)

Visual Results:



Quantitative Metrics:

While specific numerical metric like training loss were collected during testing, qualitative observations indicate a marked improvement in both the depth of reasoning and the clarity of final answers.

RESULTS(SCREEN SHOT OF OVERALL RESULT)

Train/Loss:

Description: This graph tracks the model's loss over training steps.

Observation: The loss starts around 1.9 and steadily decreases, reaching near 0.3 by step 60. This suggests that the model is learning effectively, with no sudden spikes indicating instability.

Step	Training Loss
10	1.913800
20	1.468400
30	1.408800
40	1.315300
50	1.351000
60	1.321200

Text Dataset

1	3.565200
5	2.854900
10	1.404200
15	1.269100
20	1.188500
25	1.198700
30	1.105200

Image Dataset

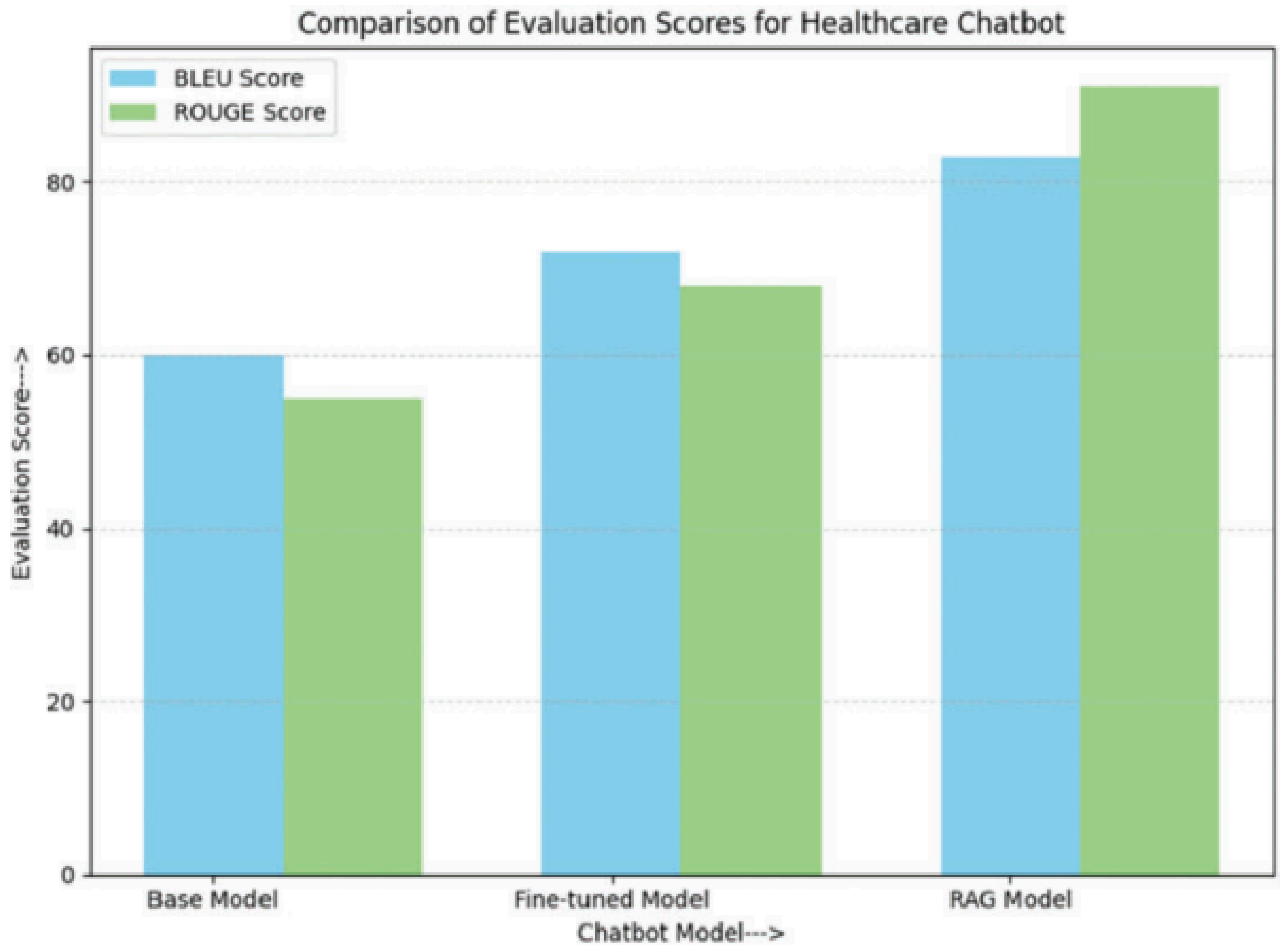


Figure 4: Graphical representation of evaluation scores

CONCLUSION

The development of AI Doctor presents a robust solution for real-time, multi-modal medical diagnostics by integrating text, image, and voice inputs into a unified, explainable pipeline. Leveraging fine-tuned transformer-based models such as DeepSeek, LLaMA, and Whisper, the system delivers accurate, clinician-friendly outputs within sub-second latency. The intuitive Gradio interface enhances usability for both medical professionals and patients, supporting accessible and informed decision-making.

Future enhancements planned for the system include:

1. Integration of live diagnostic modalities such as ultrasound and histopathology.
2. Support for multilingual inputs and culturally adapted prompts.
3. Deployment as a mobile application for low resource clinical settings.

These enhancements will further strengthen the system's clinical utility, deployment scalability, and impact across real-world healthcare and telemedicine environments.

Thank You