Curabot

Overview

Introduction

CuraBot is a **multimodal AI** medical assistant integrating text, audio, and image analysis. It supports clinical tasks like **medical query answering, X-ray generation, and brain tumor segmentation**.
Designed for efficiency, it runs on limited resources while delivering advanced diagnostic support.

Curabot system architecture

01 **Audio Input:** Processed by Whisper to convert speech into text for downstream reasoning.

02 **Text Input:** Direct input fed to Qwen combined with RAG for enhanced medical knowledge retrieval.

03 **Image generation:** Stable Diffusion produces clinical images based on textual symptom data.

04 **Image segmentation:** X-ray/CT scans analyzed with YOLO for tumor detection and segmentation.

Chatbot finetuning

Fine-Tuning Dataset

Utilizes lavita/ChatDoctor-HealthCareMagic-100k, a clean and domain-specific medical question-answer dataset to enhance accuracy in healthcare topics.

Dataset link

Model Selection

Employs Qwen/Qwen2.5-3B-Instruct quantized to 4-bit, balancing high performance with computational efficiency critical for deployment on limited hardware.

Training Approach

Uses LoRA with rank 8 for low-resource fine-tuning, enabling efficient updates to the model without retraining from scratch.

Code link

Curabot deployment

- Deployed the CuraBot interface using Flask web framework
- Integrated and hosted three core models (chatbot, image generation, tumor segmentation) using the Modals platform
- Ensures smooth API-based communication between the frontend and models
- Designed for easy access and scalability in real-world healthcare settings

Audio to text using whisper

Model Input

Incorporates Whisper basic model, optimized for 16 kHz sampled audio inputs, providing sufficient accuracy for medical speech transcription.

Design Advantages

Lightweight, yet sufficiently precise, Whisper enables real-time transcription of spoken medical queries, facilitating hands-free interaction in clinical environments.

Retrieval-Augmented Generation for Enhanced Medical QA

Vector Database & Dataset

FAISS vector store utilized for efficient retrieval from 270k Q&A samples in the DSWF/ai_medical_chatbot_train dataset.

Dataset link

Embedding and Retrieval

Uses all-MiniLM-L6-v2 embedding model to represent queries, retrieving top 10 relevant examples that contextualize and improve answer accuracy.

Chunking Strategy

Entire Q&A pairs act as retrieval chunks, maintaining semantic integrity to maximize relevance and reduce retrieval noise.

Code link

Preparing data for Diffusion

Dataset source

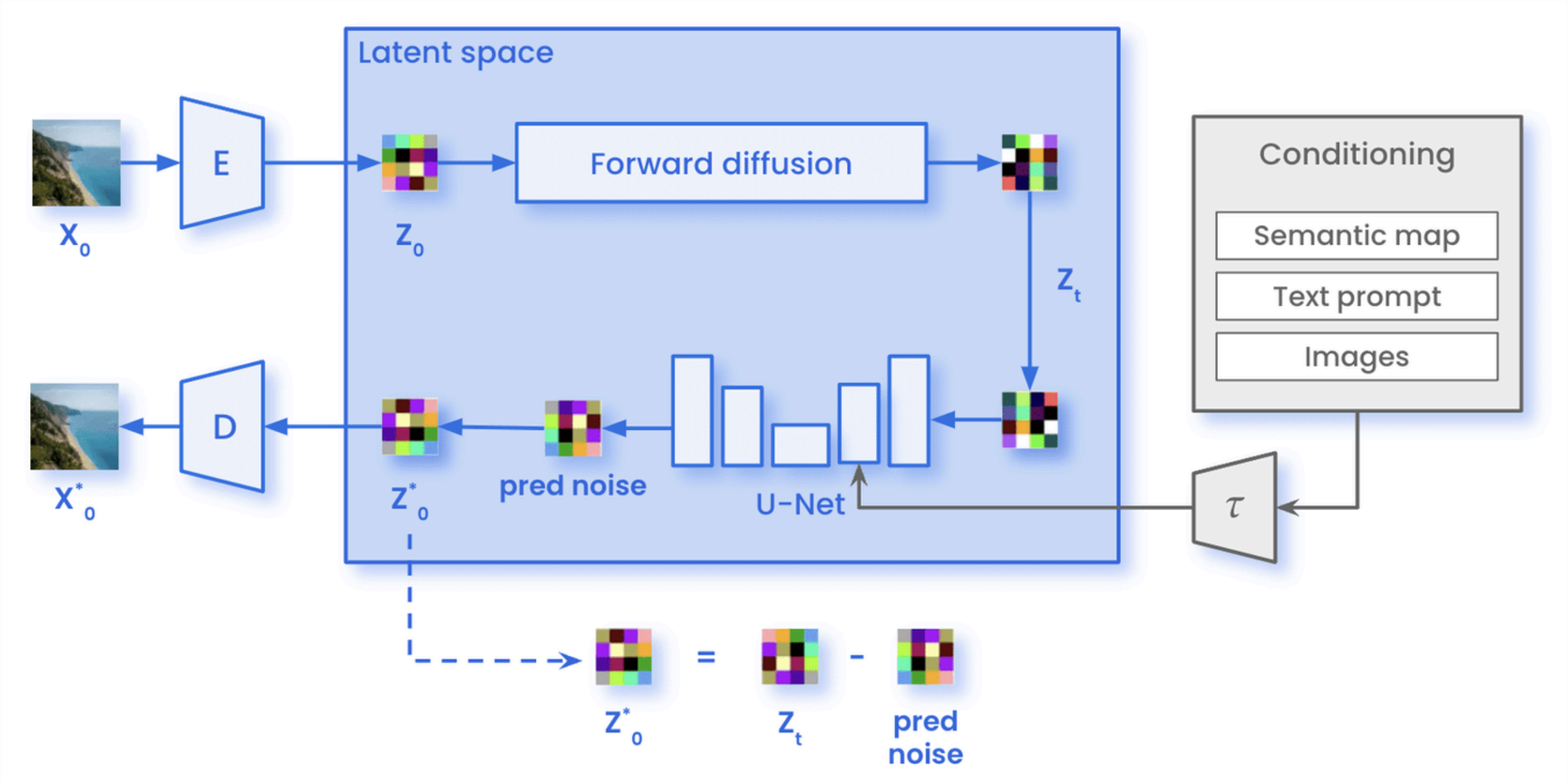
Utilized captioning_dataset_CLEF, containing X-ray, CT, sonar, and MRI images with corresponding radiology reports.

Processing captions

Applied Medical_Doctor_AI_LoRA-Mistral-7B-Instruct_FullModel from Hugging Face to convert expert-written reports into patient-spoken symptom descriptions, to create more valuable and patient-centered dataset .

Dataset link

Stable diffusion architecture



Fine tuning Stable diffusion by LoRA

Training pipeline

- Load dataset (from Hugging Face or local)
- Preprocess images and symptom prompts
- Fine-tune LoRA weights with diffusion-based loss
- Periodic image validation & checkpoint saving

Fine tuning code

LoRA Fine-Tuning Features

- Target Modules: to_q, to_k, to_v in UNet.
- Rank: Controls size of low-rank matrices (default = 4)
- Initialization: Gaussian for stability
- Efficiency: Trains only LoRA layers → low compute, high impact.

Training code

Brain Tumor Segmentation by yolo

Model and dataset

YOLO model fine-tuned on brain-tumor-yzzav dataset to detect tumor locations and classify tumor types accurately.

Performance metrics

Achieves mAP@50 of 0.992 and mAP@50-95 of 0.798, demonstrating high precision and reliable segmentation for clinical application.

Code link

Curabot deployment

- Deployed the CuraBot interface using Flask web framework
- Integrated and hosted three core models (chatbot, image generation, tumor segmentation) using the Modals platform
- Ensures smooth API-based communication between the frontend and models
- Designed for easy access and scalability in real-world healthcare settings

THANK YOU!