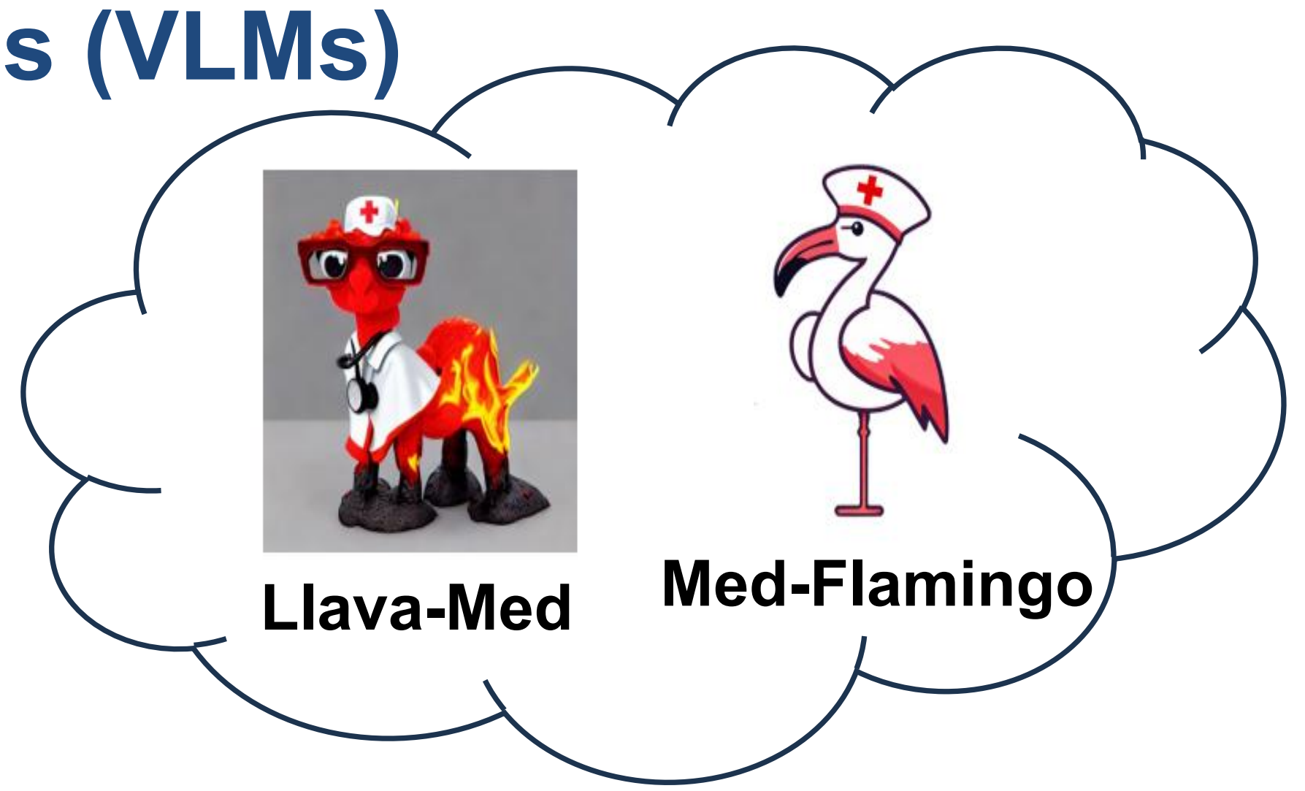
SilVar-Med: A Speech-Driven Visual Language Model for Explainable Abnormality Detection in Medical Imaging

Tan Hanh Pham[1], *Chris Ngo*[2], *Trong Duong Bui*[3], *Minh Luu Quang*[4], *Tan Huong Pham*[5], *Truong Son Hy*[6]
[1]*Harvard Medical School, USA*, [2]*Knovel Engineering Lab, Singapore*, [3]*Vietnam Military Medical University*, [4]*108 Military Central Hospital, Vietnam*,
[5]*Can Tho University of Medicine and Pharmacy*, [6]*University of Alabama at Birmingham, USA*

Motivation:

Medical vision language models (VLMs)

Existing medical VLMs: Med-Flamingo, Llava-Med, MiniGPT-Med, MedBLIP, …



- Most of existing medical VLMs do not provide reasoning behind medical image diagnosis.
- They are limited by text+image instruction models.

Potential problem

Figure 1: Surgery (Harvard Medical School)

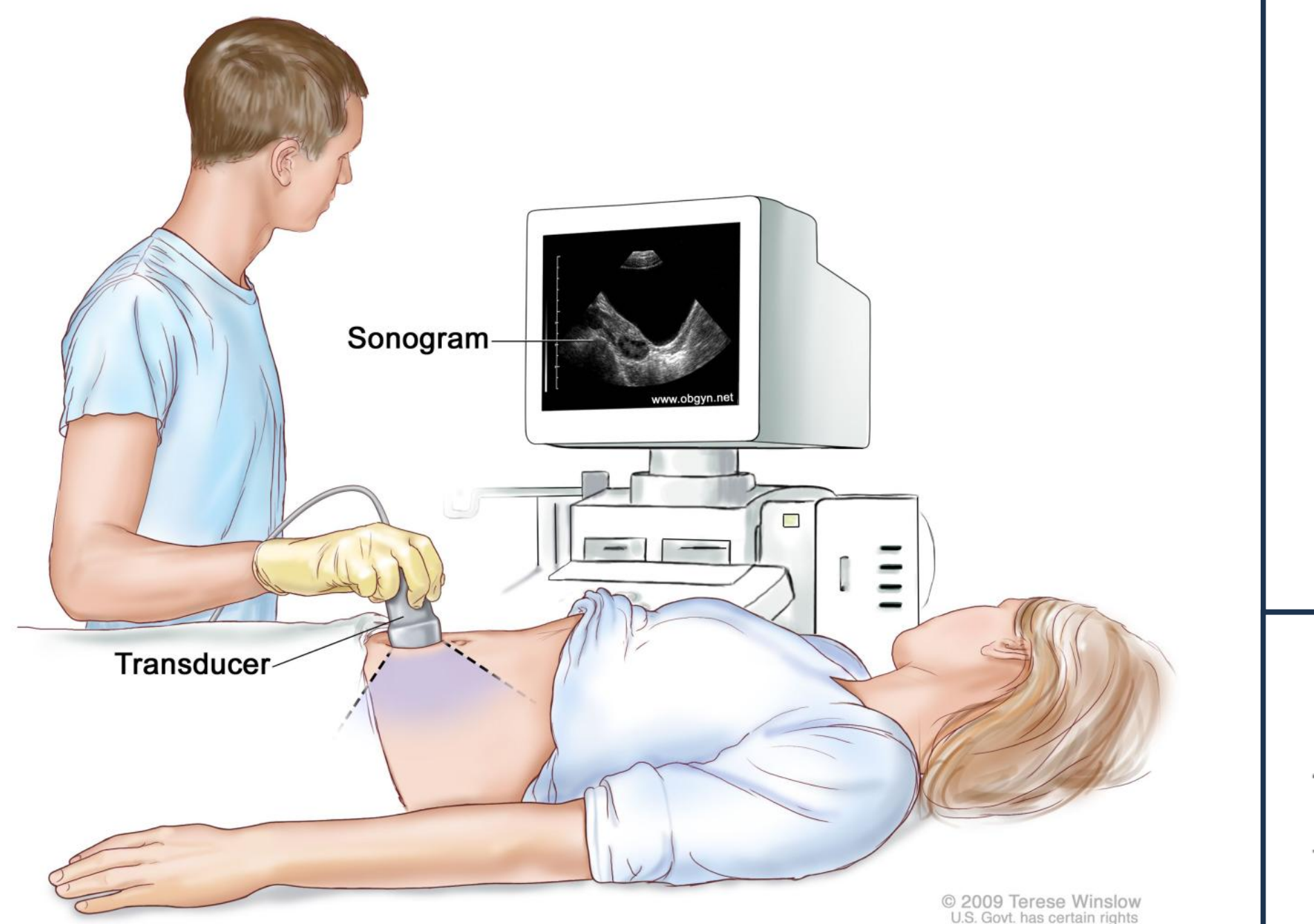


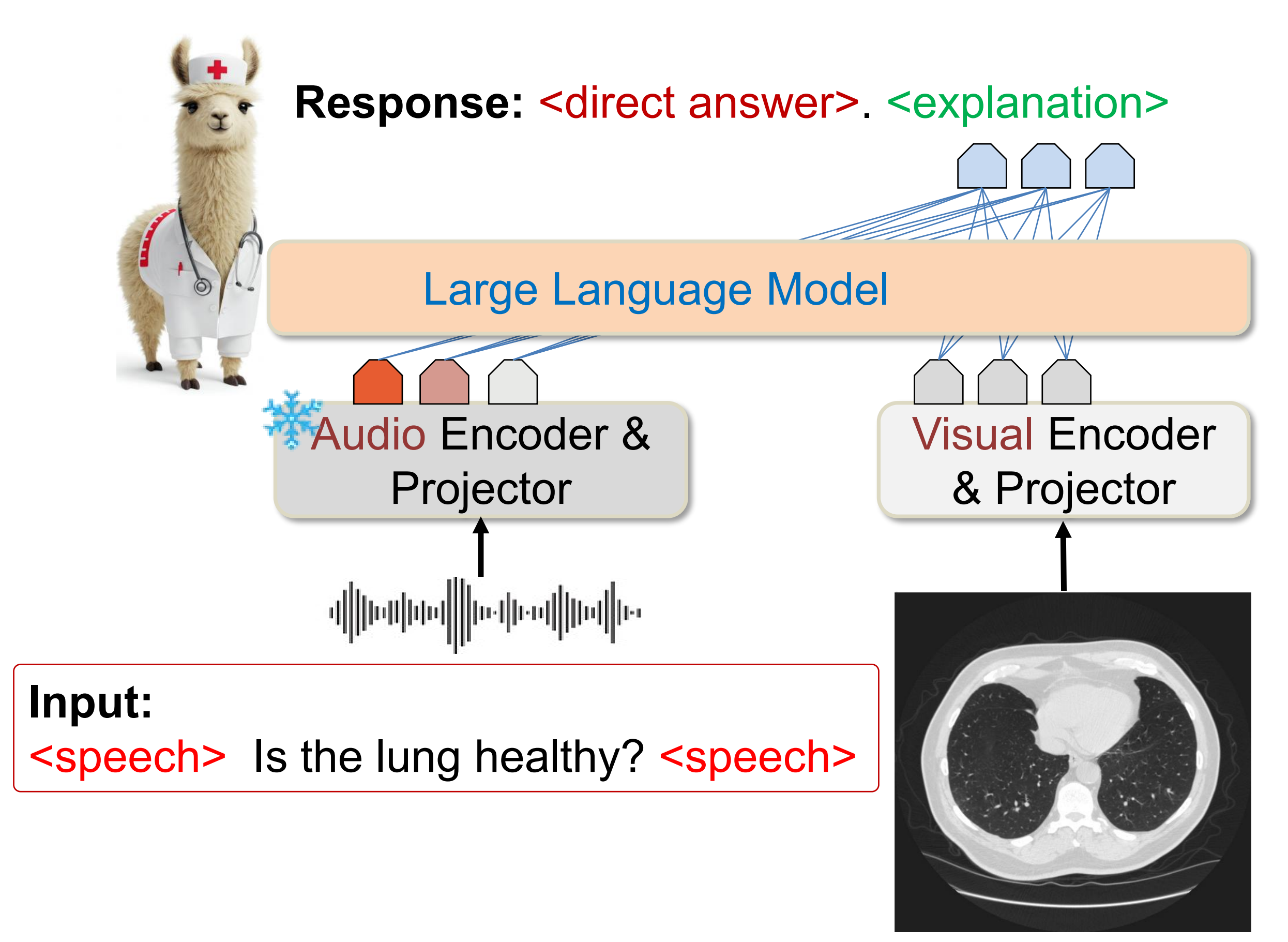
Figure 2: Abdominal ultrasound (National Cancer Institute)

➡ There is a potential for speech-instruction medical VLMs (audio+image) in scenarios such as surgery or text-based interaction is often impractical for physicians.

Contribution

- **Method**: We propose SilVar-Med, a ***speech-driven medical VLM*** that enables intuitive human-machine interaction in healthcare.
- **Dataset benchmark**: a ***reasoning*** dataset for abnormality detection
- **Evaluation metrics**: ***Model's reasoning*** abilities and human evaluation

Methodology



Response: <direct answer>. <explanation>

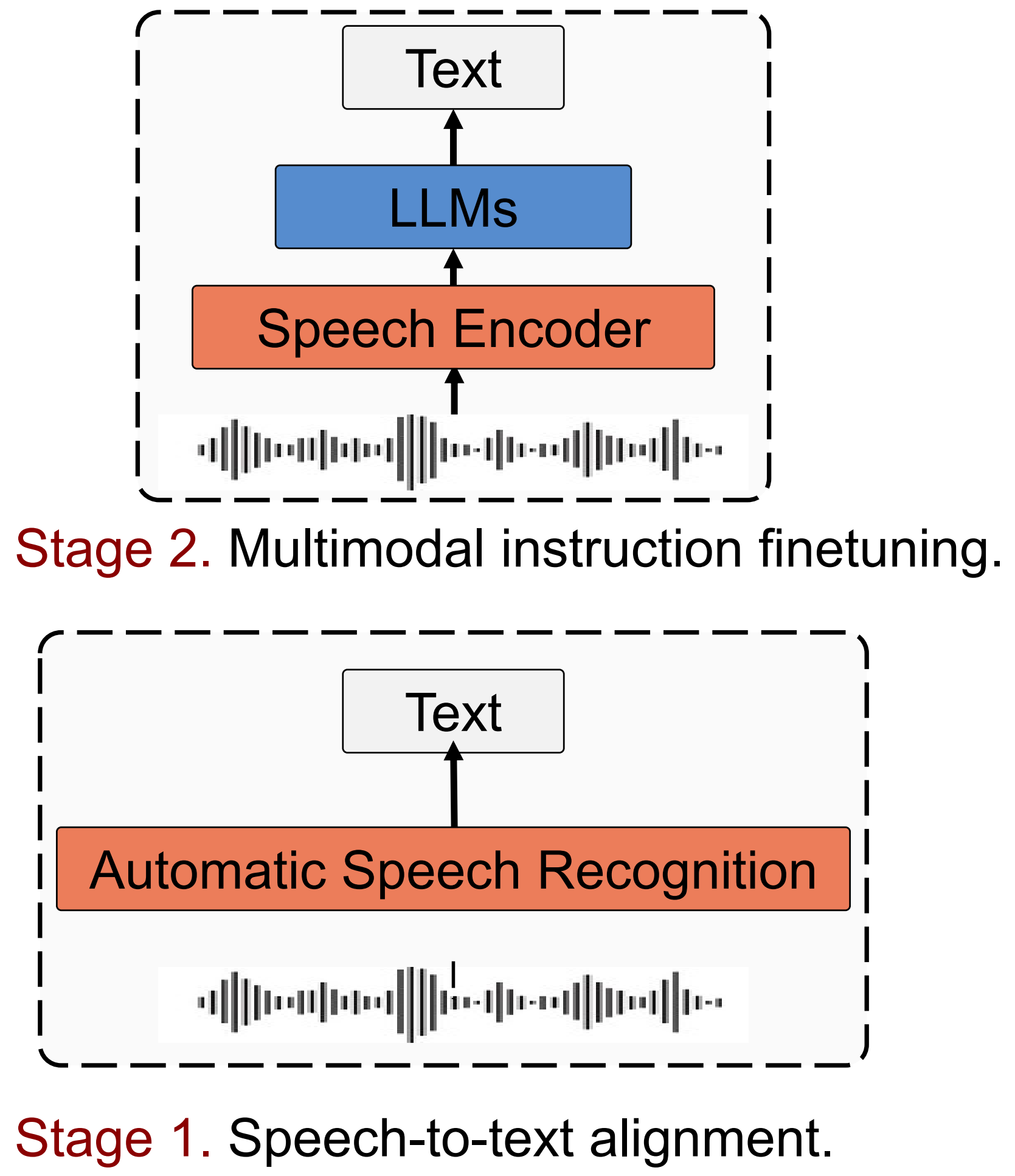
Large Language Model

Audio Encoder & Projector

Visual Encoder & Projector

Input:
<speech> Is the lung healthy? <speech>

Figure 3: Speech-Driven Medical VLM for Explainable Abnormality Detection in Medical Imaging.



Text
LLMs
Speech Encoder

Stage 2. Multimodal instruction finetuning.

Text
Automatic Speech Recognition

Stage 1. Speech-to-text alignment.

Figure 4: Training pipeline.

LLM Llama 3.1-8B | Audio encoder Whisper Tiny | Visual Encoder PubMedCLIP

Datasets

Dataset	Train	Validation	Test
VQA-RAD	1.7k	None	451
SLAKE (English)	4.9k	1k	1k
VQA-Med 2019	12.7k	2k	500
SilVar-Med (ours)	716	-	150

Table 1: Training dataset

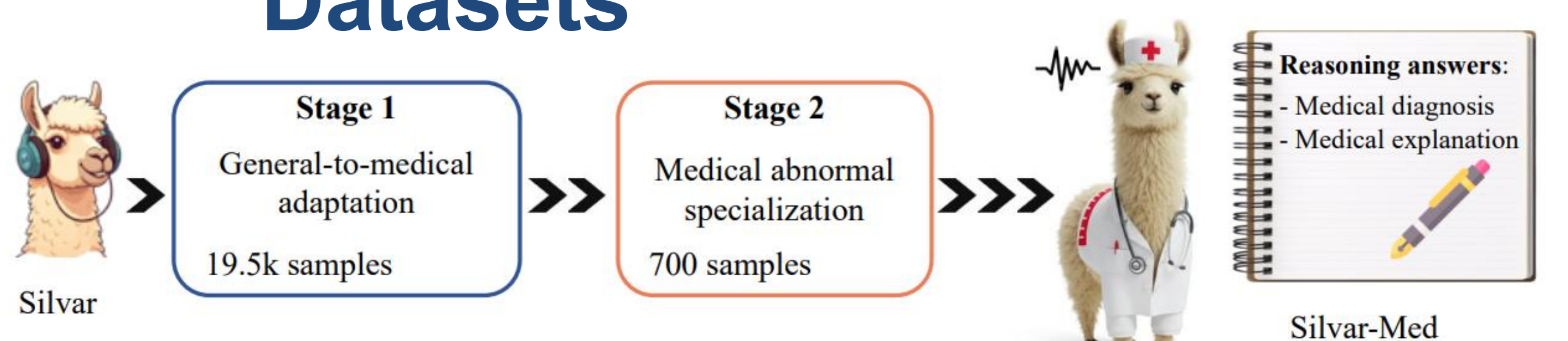
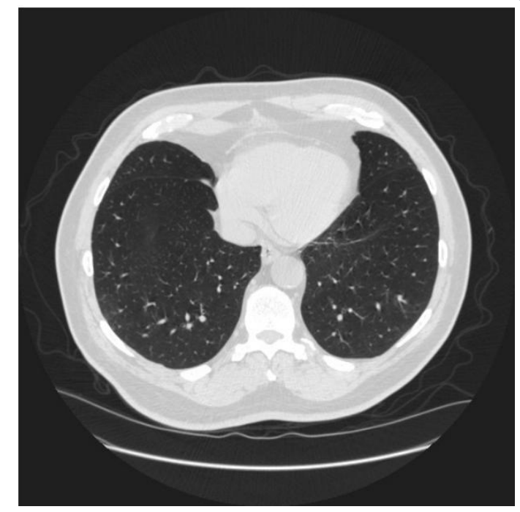
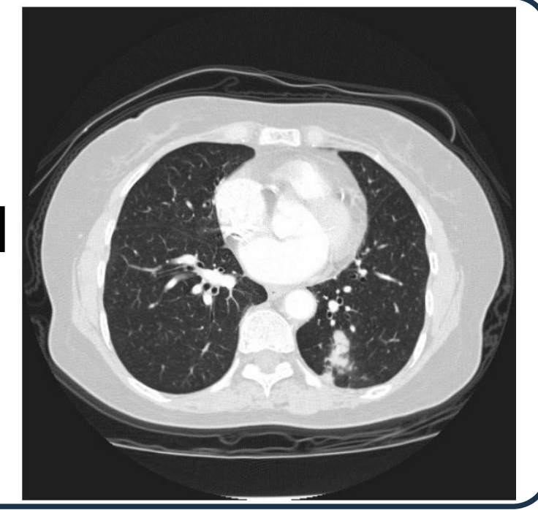


Figure 5: Adaptive training pipeline

Sample 1: No abnormality detection
Question: Is the lung healthy?
Answer: <answer>The lung appears healthy based on the provided imaging.<answer> <explanation> There are no visible signs of abnormalities such as lesions, fluid accumulation, or structural changes. The lung fields are clear, indicating normal function and absence of disease. <explanation>



Sample 2: Abnormality detection
Question: What diseases are included in the picture?
Answer: <answer> The image indicates the presence of lung cancer, characterized by abnormal growths in the lung tissue. <answer> <explanation> The CT scan shows potential lesions or nodules that may suggest malignancy. <explanation>



Experimental Results

Automatic speech recognition

Models	WER		CER	
	train	test	train	test
Whisper Tiny	2.01	2.67	2.01	2.99
Whisper Small	2.02	4.57	1.59	3.50

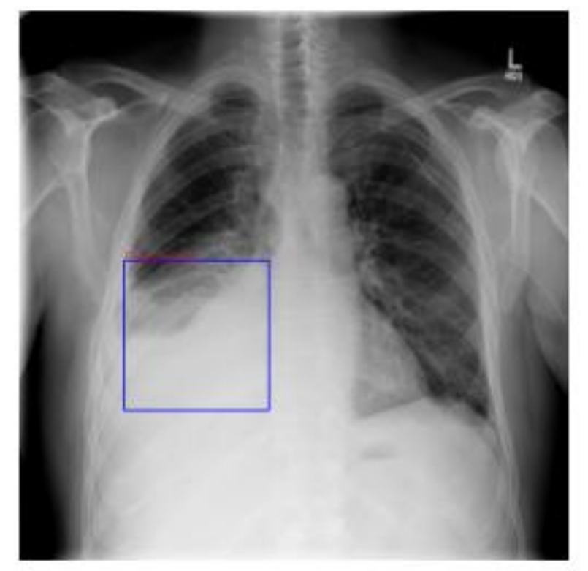
Table 2: Automatic speech recognition.

Reasoning evaluation

Reasoning accuracy	Exp 1	Exp 2	Exp 3	GPT4o	Gemini
SilVar-Med with the langue module of Llama 3.1 8B					
Completely Incorrect	11	6	13	39.00	22.00
Significantly Incorrect	28	30	33	9.67	23.67
Partially Correct	13	15	28	39.67	54.00
Fully Correct	96	97	74	59.67	48.33

Table 3: Reasoning evaluation (Exp denotes expert).

Example 1: Abnormality detection.
🔊 Does the lung look abnormal?



Ground truth: Yes, the lung appears abnormal as there are irregularities in the lung fields that may indicate potential pathology. The presence of unusual opacities or changes in the lung structure can suggest conditions such as infection, inflammation, or other lung diseases.

SilVar-Med (Llama 3.1): Yes, the lung appears abnormal due to the presence of irregularities in the lung fields, which may indicate potential pathology such as infection, inflammation, or other lung conditions. The overall lung structure shows signs of asymmetry or density changes that deviate from normal anatomy. These findings warrant further investigation to determine the underlying cause.

GPT-4o mini: The lung appears to be within normal limits based on the provided X-ray image. There are no obvious signs of abnormalities such as masses, infiltrates, or significant opacities. The lung fields are clear, and the vascular markings are normal, indicating healthy lung tissue. However, a definitive assessment should be made by a qualified radiologist for a comprehensive evaluation.

Gemini 1.5 Flash: Based on the provided chest X-ray, the lungs appear normal. There is no visible evidence of infiltrates, masses, or pneumothorax. The lung fields demonstrate normal aeration and the heart size appears within normal limits. Further evaluation may be necessary depending on the clinical presentation.

Conclusion

- **Model**: We propose SilVar-Med, a **speech-driven medical VLM** that enables intuitive human-machine interaction in healthcare.
- **Dataset**: Reasoning dataset and extensive experiments for abnormality detection.
- **Evaluation metrics**: Model reasoning using LLM-as-judge and human evaluation,.. There is a gap between human and LLM-as-judge evaluation.
- **Application**: Verbal communication with VLM (e.g. surgery, abdominal ultrasound, …)

Project page

Silvar-Med Paper