EchoVLM: Report Summarizer

Stanford CS 195 - Dr. Ehsan Adeli

Syed Ashal Ali
Department of Computer Science
Stanford University
ashal@stanford.edu

1 Key Information

- Mentor: Dr. Chieh-Ju Chao
- Affiliation: Stanford AI Lab, PAC

2 Introduction and Previous Work

The previous approach to EchoGPT is built on an open-source base—LLaMA-2—adapted through Quantized Low-Rank Adaptation (QLoRA). This approach fine-tunes only small, low-rank adapter modules, allowing efficient training, and enabling EchoGPT to capture the style of echocardiography reporting. During training, large numbers of "Findings → Impression" pairs are fed into the model, together with instructions (e.g., "You are a knowledgeable cardiologist..."), so it learns exactly which items to highlight: pertinent measurements (like ejection fraction and valve gradients), clinically significant morphological details, and key conclusions. At inference time, new "Findings" are appended to a short prompt, and EchoGPT generates a streamlined "Impression" that mirrors real-world cardiologist practice.

Quantitatively, EchoGPT demonstrates substantial gains over simpler methods or baseline LLMs, consistently outscoring them on automated natural language metrics (BLEU, METEOR, ROUGE-L, BERTScore), and also showing fewer factual errors on a specialized domain metric called RadGraph F1. Of particular note, when cardiologists evaluated these generated summaries head-to-head with the physician-written originals, EchoGPT's versions were deemed equivalently correct and clinically useful, while often being more concise. This success suggests that the QLoRA fine-tuning strategy can be highly effective, particularly when numeric precision matters. Overall, EchoGPT offers a robust blueprint for domain-specific text summarization models in healthcare, paving the way for more advanced "co-pilot" systems to reduce physician documentation workloads while safeguarding the clarity and accuracy of patient reports [1].

3 Current Motivation

As per Figure 1, we can see that the previous EchoGPT approach taught us that fine-tuning may work at one institution, but there are limitations in generalization. In order to solve the generalization problem, we explored implementing an agentic architecture in order to be able to produce cardiology reports that are accurate and in the correct style of cardiology reporting across any institution and echocardiography dataset.

AI agents are defined as systems that can "can plan and execute interactions in open-ended environments, such as making phone calls or buying online goods." [2] and have "the ability to act without human intervention or other systems." This autonomous behavior of AI agents is why we believe that they would be a prudent approach to the generalization problem. [3].

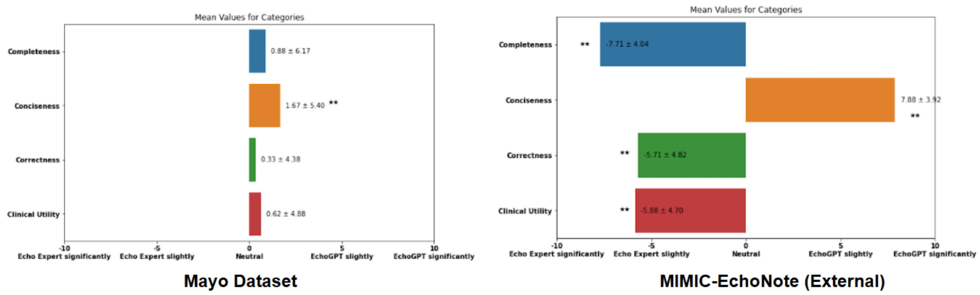

Figure 1: This figure displays preference ratings of the generated "Impression" evaluated by cardiologists. The "Impression" from ground truth report versus EchoGPT were blindly across the four metrics corresponding to the bars in the figure. EchoGPT does not perform as well on MIMIC than it does on the MAYO dataset (the dataset it was finetuned on) [1].

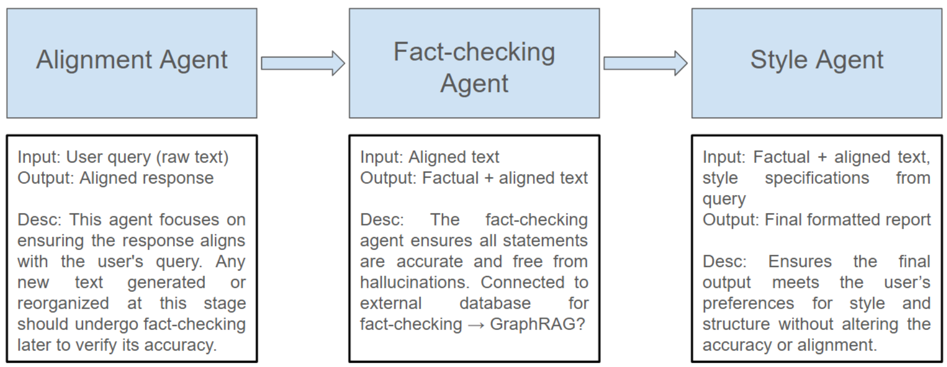

Figure 2: This figure presents a brief overview of the three different agents that are currently part of the report summary generation pipeline. The pipeline includes Alignment, Fact-checking, and Style agents, which are all based on GPT-4.

4 Approach

This quarter, I employed a multi-agent pipeline to transform the text data part of the MIMIC dataset into a concise and accurate "Impression" in the style of an ECG report. This pipeline is organized into three discrete components: Alignment, Fact-Checking, and Style Agents—each performing a specialized function while passing its output to the next stage.

In the first step, the Alignment Agent receives the user's unstructured echocardiographic impression and produces a short medical summary that aligns with the query or context. This stage preserves the core meaning of the original text but defers in-depth verification to later modules. Next, the Fact-Checking Agent compares the aligned summary against the original raw text (before impression), correcting any factual inconsistencies or speculative elements. Here, the system explicitly cross-references the new summary with the original source to minimize hallucinations and ensure the final content remains medically faithful. Such checks could, in principle, also query external approaches that members of the lab have been working on (e.g., GraphRAG) for additional evidence, although my present work solely focuses on direct text-to-text validation.

Finally, the Style Agent takes the fact-checked output and reformats it according to standard cardiology reporting conventions. Guided by sample templates, it adjusts structure, terminology, and stylistic presentation without altering substantive medical content.

A summarized version of the pipeline is displayed in Figure 2.

Each agent is implemented as a separate chain of prompts (via ChatOpenAI's GPT-4 model) within a Python framework, allowing for modular development and testing. I load and process multiple rows of training data, routing each impression sequentially through these three specialized agents.

Metric	Score
Mean BERT	80 (+/- 1.9)
Mean ROUGE-L	8 (+/- 4.4)
Mean BLEU	0.1 (+/- 0.3)
Mean METEOR	5.1 (+/- 3.5)

Table 1: This table displays the different scores that I received when testing my system on the test set and evaluating for the metrics listed.

Full Information	**Impression**	**Generated Impression**
PATIENT/TEST INFORMATION: Indication: Left ventricular function. Bacteremia/ group A strep Height: (in) 64 Weight (lb): 141 BSA (m2): 1.69 m2 BP (mm Hg): 140/77 HR (bpm): 120 Status: Inpatient Date/Time: [**2127-10-23**] at 09:15 Test: Portable TTE (Congenital, complete) Doppler: Full Doppler and color Doppler Contrast: Saline Technical Quality: Adequate INTERPRETATION: Findings: LEFT ATRIUM: Mild LA enlargement. RIGHT ATRIUM/INTERATRIAL SEPTUM: Normal RA size. Aneurysmal interatrial septum. PFO is present. LEFT VENTRICLE: Normal LV wall thickness, cavity size, and systolic function (LVEF>55%). Normal regional LV systolic function. No resting LVOT gradient. RIGHT VENTRICLE: Normal RV chamber size and free wall motion. AORTA: Normal aortic root diameter. Normal ascending aorta diameter. Normal aortic arch diameter. AORTIC VALVE: Normal aortic valve leaflets (3). No AS. No AR. No masses or vegetations on aortic valve. MITRAL VALVE: Normal mitral valve leaflets. No MVP. No mass or vegetation on mitral valve. Mild (1+) MR. TRICUSPID VALVE: Normal tricuspid valve leaflets. No mass or vegetation on tricuspid valve. Mild [1+] TR. Normal PA systolic pressure. PULMONIC VALVE/PULMONARY ARTERY: Normal pulmonic valve leaflets with physiologic PR. PERICARDIUM: Trivial/physiologic pericardial effusion. GENERAL COMMENTS: Resting tachycardia (HR>100bpm). Left pleural effusion. Conclusions: The left atrium is mildly dilated. The interatrial septum is aneurysmal, and there is right-to-left flow of saline contrast, indicating a patent foramen ovale. Left ventricular wall thickness, cavity size, and systolic function are normal (LVEF>55%). Regional left ventricular wall motion is normal. Right ventricular chamber size and free wall motion are normal. The aortic valve leaflets (3) appear structurally normal with good leaflet excursion and no aortic regurgitation. No masses or vegetations are seen on the aortic valve. The mitral valve leaflets are structurally normal. There is no mitral valve prolapse. No mass or vegetation is seen on the mitral valve. Mild (1+) mitral regurgitation is seen. The estimated pulmonary artery systolic pressure is normal. There is a trivial/physiologic pericardial effusion.	No evidence of endocarditis. Normal global and regional biventricular systolic function. Patent foramen ovale.	Echocardiography reveals an overall normal condition of the patient\\'s heart with minor abnormalities. In relation to the heart\\'s chambers, there is a minor enlargement of the left atrium. An unusual formation is noted in the form of an aneurysmal interatrial septum accompanied by a patent foramen ovale. Slight irregularities are observed in the heart\\'s valves such as mild (1+) mitral regurgitation and trivial tricuspid regurgitation. Minor pericardial effusion and left pleural effusion are also apparent. Noteworthy is that the patient is undergoing inpatient monitoring for left ventricular function due to bacteremia/group A strep, and these observations compare with the expected readings from such a condition. Based on current AHA endocarditis prophylaxis recommendations, these echo findings indicate a cautious evaluation. Clinical decisions regarding the need for further actions should be thoroughly considered and based on both clinical and echocardiographic data. Comparing with previous records, the current findings are mainly consistent, maintaining a similar condition.

Figure 3: This figure displays an observation from my generated results dataset. The "full information," "impression", and "generated impression" are visible. The differences between the ground truth impression and the generated impression are described in the Analysis section.

The asynchronous design improves throughput, and the final output—now condensed, vetted, and formatted—can be directly appended to a physician's workflow as a draft for further clinical review. This staged "agentic" approach thus provides clearer accountability for each transformation of the text, ensuring alignment with user needs, factual correctness, and appropriate clinical style.

5 Experiments and Discussion

I evaluated the quality of the system's echocardiography summaries using BLEU, METEOR, ROUGE-L, and BERT metrics on the test set. After loading the dataset, each generated output was compared to the reference summary (the original "response"). I computed BLEU, METEOR, and ROUGE-L via standard NLTK-based routines to gauge n-gram overlaps, and used BERT score to measure semantic similarity.

The final scores shown in Table 1 (e.g., BLEU at 0.1 ± 0.3) reveal the model's exact-token overlap (low BLEU, METEOR, ROUGE-L) versus its higher semantic alignment in BERT scores. This highlights that the model is generating a summary that has a similar meaning to the ground truth impression, but does not follow the same style of reporting.

6 Analysis

As seen in Figure 3, the generated impression is much more verbose than the ground truth impression. The reference ("Impression") is sharply focused on the most clinically pivotal findings—namely, ruling out endocarditis, confirming normal global and regional systolic function, and noting the patent foramen ovale. By contrast, the model-generated content delves into additional details, including mild regurgitations and minor pericardial effusion, as well as contextual mentions of Group A strep bacteremia. Even though the overall information captured by the ground truth and generated impressions are similar, the style of cardiology reporting is much more consistent with the ground truth as opposed to the generated impression.

7 Conclusion and Future Direction

In summary, I introduced a multi-agent pipeline—featuring Alignment, Fact-Checking, and Style Agents—to address the generalization limitations of purely fine-tuned echocardiography report summarizers. By breaking down the summarization task into distinct stages, the system ensures that each transformation (aligning content, verifying facts, and enforcing stylistic consistency) is traceable and tailored to clinical requirements. While initial evaluations show that the generated impressions maintain semantic fidelity, the style and level of detail still differ from concise human-written conclusions. Future work is centered upon improving the results. Currently (and over the break), I am iterating upon the prompts being used by the architecture and sending the intermediate results (the output from every agent) to Dr. Chao in an attempt to improve the quality of results of the system.

References

[1] Chieh-Ju et al. Chao. Echogpt: A large language model for echocardiography report summarization. *medRxiv*, 2024.

[2] Alan et al. Chan. Infrastructure for ai agents, 2025.

[3] Ira Rudowsky. Intelligent agents by i. s. rudowsky INTELLIGENT AGENTS, 2008.