SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

RadFeedback: Generate Individualized Trainee Feedback using Large Language Models for Analysis of Preliminary-to-Final Radiology Report Changes

Piyush Goyal, DO, University of Utah; Rachel Dou, MD; Sarah Teruya, MD; William Auffermann, MD, PhD

Briefly describe what attendees will see during your demo.
Attendees will see a graphical user interface for uploading datasets consisting of paired trainee-preliminary and attending-finalized radiology reports. The program will generate an individualized LLM-based feedback summary for each trainee. Users can review the output, enter their comments into the program, and those comments can be used to guide future iterations of the LLM-prompt used on the backend to generate the trainee feedback.

Problem & Motivation: What clinical or operational problem does your MVP address?
Radiology trainees learn substantially from comparing preliminary reports with attending-finalized reports, but manual review is time-consuming and often obscured by stylistic edits. We sought to build a LLM-powered workflow that converts paired resident-attending report differences into actionable feedback that identifies recurring knowledge gaps and opportunities for deliberate practice.

What You Built: Describe your MVP, tool, or workflow in plain language
We developed a secure pipeline that 1) paired each trainee preliminary report with its attending finalized report, 2) removed personal health information (PHI) from the dataset, and 3) analyzed the paired dataset in an isolated workspace (ChatGPT Enterprise) for actionable report changes (Figure 1). Inputs included preliminary text generated by a radiology trainee, final text attested by an attending radiologist, and study type. A prompt was iterated to direct the LLM to focus on clinically significant changes – changed impressions or missed findings. Style-only changes were excluded unless the same pattern was applied by more than one attending. Outputs for each trainee were aggregated by study type and prioritized by clinical impact to generate a summarized feedback document. Trainee perceptions were surveyed before and after receiving the feedback.

Tech Stack & Development Approach: What tools, frameworks, or methods did you use to build it?
Python to format and organize the extracted data. LLM (ChatGPT) to analyze the data and generate the output (individualized trainee feedback based on radiology report edits made by attending). Vibe coding to develop program mockups and the prototype graphical user interface.

Validation & Evidence: Have you done any informal testing or validation? If so, what did you find?
In a retrospective feasibility pilot study, the pipeline processed 6,468 paired preliminary-final reports over the course of 3 months from six trainees in their final year of training who agreed to participate in the study. Textual changes were present in 4,039 reports (62.4%), demonstrating that unfiltered report comparison

would create a high feedback burden. The workflow generated individualized feedback that was aimed at summarizing clinically meaningful patterns of report changes, highlight recurring knowledge gaps with linked peer-reviewed resources to address these gaps, and produced targeted checklists to refine search pattern. On average trainees responded spending 60 minutes per week reviewing their reports. Prior to receiving the feedback, trainees had neutral to favorable perception of the LLM-feedback in accurately identifying knowledge gaps and impacting their search patterns. After receiving the feedback, trainees' perceptions tilted toward skepticism as they noted the feedback was still heavily focused on stylistic changes (Figure 2).

Feedback You're Seeking: What specific feedback, help, or collaboration are you hoping to get from the CAIMI26 community?

We are seeking feedback on how to determine whether the LLM-prompt is truly optimized, how to increase trainee engagement with the program, and how the workflow could be broadened to maximize utility for education and further research —for example, by retrieving similar cases from another resident's experience—and whether discrepancy patterns between attendings and residents could support training data needed for vision-based training models.

Known Limitations & Honest Failures: What isn't working yet, or what have you already tried that failed?

There are limitations and failures. 1) Currently only "preliminary" and "final" reports can be compared. Extracting "draft" reports directly from PowerScribe will either require vendor support, or internal tool development to access the locally stored database. 2) Despite multiple iterations of prompt engineering, the LLM still tends to overemphasize stylistic edits. Although successive prompts have reduced this issue, further refinement is still needed to better prioritize clinically meaningful discrepancies.

Potential for Impact: If this MVP were fully developed and deployed, who would benefit and how?

This workflow can reduce the manual burden of report comparison and transform attending edits into structured formative feedback. It can help trainees rapidly identify missed findings, refine search patterns, and monitor progress over time. It also gives educators a scalable method to detect recurring discrepancies and direct targeted teaching toward study types that are challenging for multiple residents.

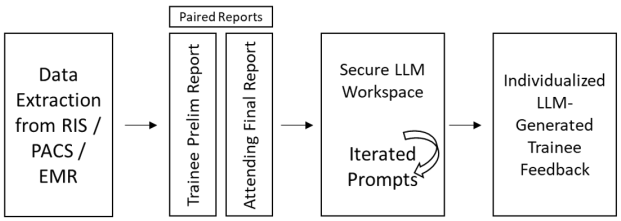


Figure 1. Workflow operationalizing review of trainee radiology reports.

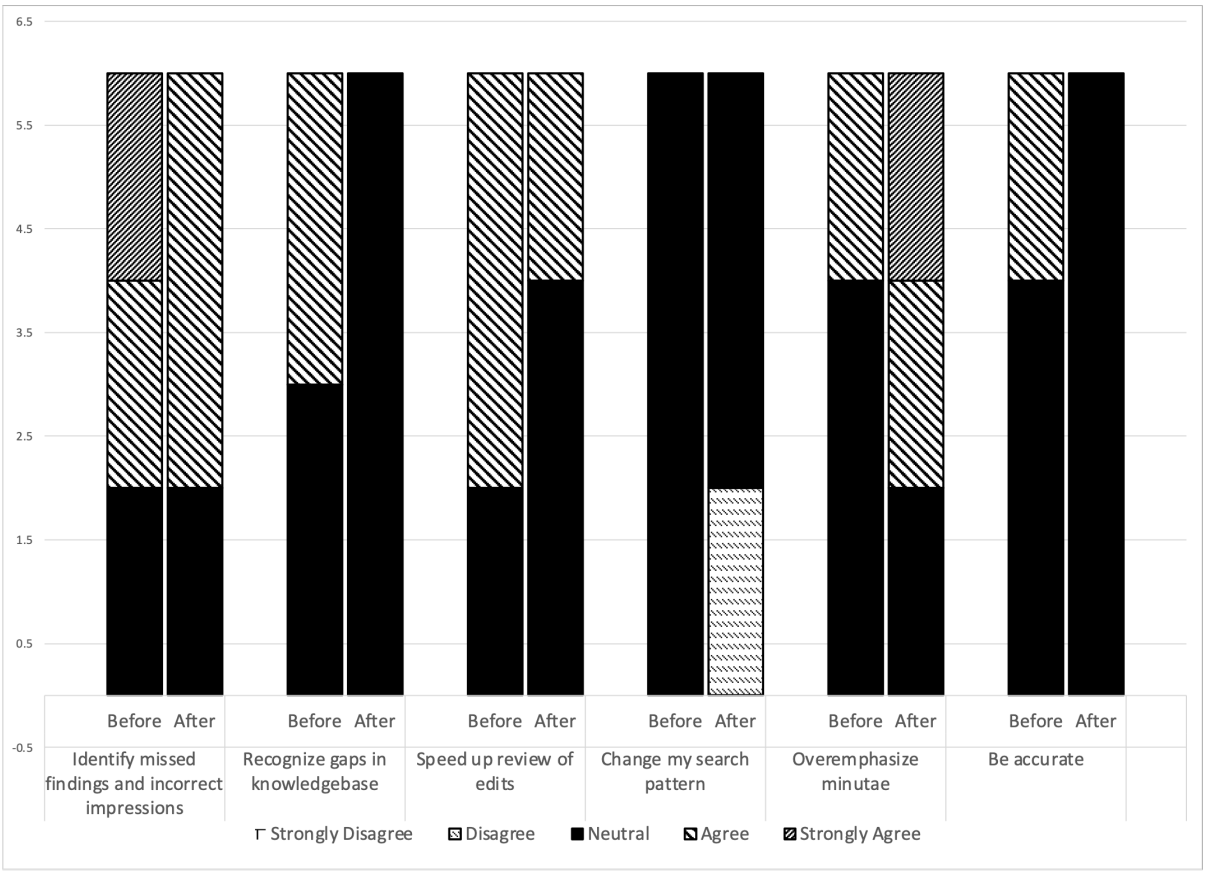
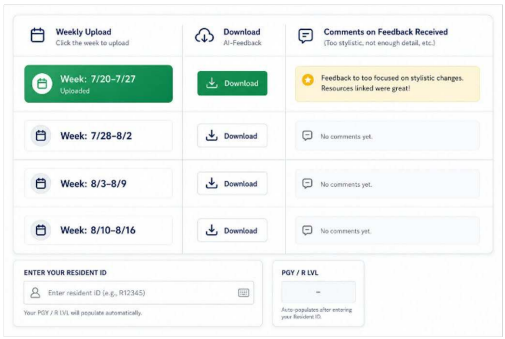
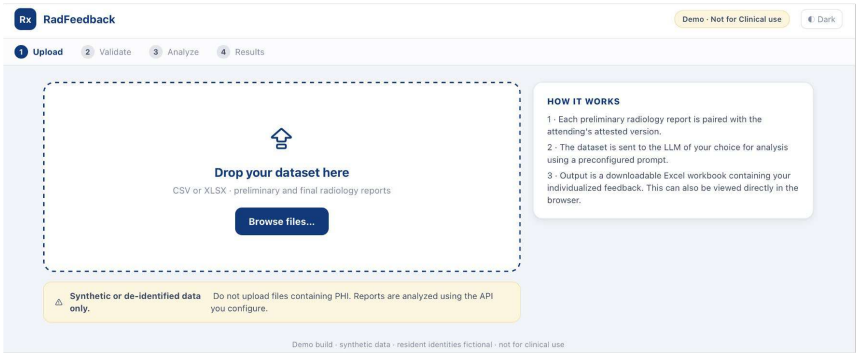


Figure 2. Survey results summarizing trainee perception of LLM-generated feedback prior to and after receiving individualized LLM-generated feedback. The survey questions listed below either started with "I expect the LLM-generated feedback to:" or "The LLM-generated feedback was able to:" depending on whether the survey was administered prior to or after providing each resident with their LLM-generated feedback.





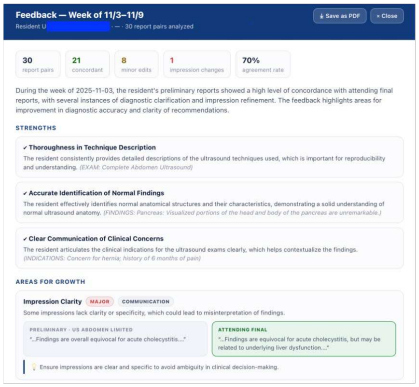


Figure 3. Screenshots from prototype program.