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

Large Language Model–Assisted Analysis of Preliminary-to-Final Radiology Report Changes for Resident Feedback

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

Introduction/Background

Radiology trainees learn substantially from comparing preliminary reports with attending-finalized reports, yet manual review is time-consuming and often scattered with relatively stylistic edits. We hypothesized that a creating a workflow where a large-language-model (LLM) periodically analyzes radiology dictation data to generate individualized trainee feedback would reduce manual review burden and improve overall radiology education.

Methods/Intervention

We developed a secured pipeline that 1) paired each trainee preliminary report with its attending finalized report, 2) removed PHI from the dataset, and 3) uploaded the paired dataset to an isolated workspace (ChatGPT Enterprise) (Figure 1). Inputs included preliminary text generated by trainee, final text attested by an attending radiologist. A basic 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 repeatedly. LLM-generated feedback for each trainee was aggregated by study type and prioritized by clinical impact to generate a summarized document. Trainee perceptions were surveyed before and after receiving the LLM-generated feedback.

Results/Outcome

In a retrospective feasibility pilot study, the pipeline processed 6,468 paired preliminary-final reports over 3 months from six graduating trainees. Textual changes were present in 4,039 reports (62.4%), demonstrating that unfiltered report comparison would create a high feedback burden. Individualized feedback workbooks were generated that were aimed at summarizing clinically meaningful patterns of report changes, recurring knowledge gaps, and targeted refinements to search pattern. On average trainees responded spending 60 minutes per week reviewing their reports' edits. Trainees initially had neutral to favorable perception of the LLM-feedback in being able to accurately identify knowledge gaps and impact their search patterns. After receiving the feedback, trainees' perceptions mildly shifted toward skepticism as they noted the feedback was heavily focused on stylistic changes (Figure 2).

Conclusion

An LLM-based report-comparison workflow can summarize edits into structured, individualized formative feedback and has the potential to identify targeted knowledge gaps. Further prompt engineering is required to further emphasize clinically meaningful changes from stylistic edits.

Statement of Impact

This LLM-based workflow can operationalize feedback by directing trainees' attention toward high-yield

changes between their preliminary reports and attending attested final reports.

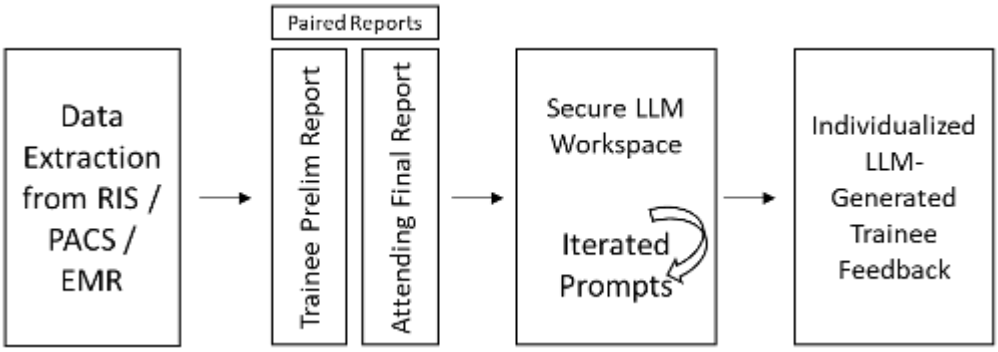

Figure 1. Workflow operationalizing review of trainee radiology reports.

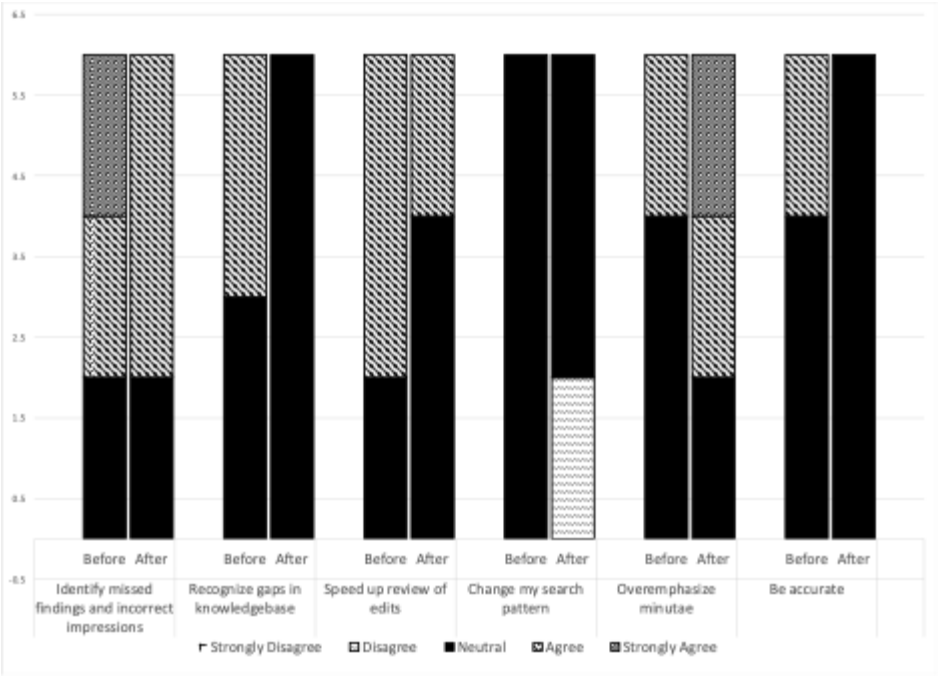

Figure 2. Survey results summarizing trainee perception of LLM-generated feedback prior to and after receiving individualized LLM-generated feedback. The survey questions listed below either started with “I expect the LLM-generated feedback to:” or “The LLM-generated feedback was able to:” depending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

Keywords

Feedback; Education; NLP; Trainee; Report Comparison; Dictation