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

Residents as a Proxy for AI Draft Reporting: LLM-based Analysis

Steven Rothenberg, MD, Thomas Jefferson University; Shahir Monsuruddin, MD; Paras Lakhani, MD; Trevor Vent, PhD; Baskaran Sundaram, MD; Adam Flanders, MD, FSIIM

Introduction/Background
No AI products are currently approved by the FDA for draft or preliminary reporting of chest x-rays (CXRs); however, many vendors and practices are starting to use these tools for drafting radiology reports. Two companies have received breakthrough device designation for interpreting CXRs and several more offer foundation models as a service or tools for developing one's own draft reporting system under a research protocol. The purpose of this study was to use large language models (LLMs) to explore patterns that emerge when provided with a draft report using residents as a proxy for AI-based methods.

Methods/Intervention
This retrospective observational study was determined to be exempt by the IRB. CXRs interpreted between May 1, 2026 and June 30, 2026 were included. Preliminary reports, final reports, and examination metadata were extracted from the radiology reporting system. Findings and impression sections were parsed and processed using two separate LLM pipelines based on with qwen2.5-14b to extract findings and flag reports for potential major discrepancies. Reports were classified as unchanged when preliminary and final reports match exactly after normalization for changes in punctuation and white space. Revision and discrepancy rates were compared by daytime versus overnight interpretations and weekend versus weekday interpretations.

Results/Outcome
A total of 7,242 CXRs were included of which 68.6% (4,968/7,242) were unchanged, 4.6% (334/7,242) were flagged for potential major discrepancy. Overall, 11.5% (833/7,242) were flagged or had more than 10% of the report text changed. 23.1% (15/65) of shifts with greater than 30 CXRs demonstrated unchanged report rates greater than 90% suggestive of automation bias. Fewer reports were changed on weekend compared to weekdays (OR 0.773, 95% CI 0.687 – 0.870).

Conclusion
Academic radiologists change fewer resident reports at the end of high-volume shifts and on weekends. Therefore, these times of reporting are at potentially higher risk for automation bias with AI draft reporting.

Statement of Impact
Understanding resident reporting workflows may inform necessary guardrails and tracking metrics to minimize automation bias during implementation of AI-based draft reporting for CXR interpretation.

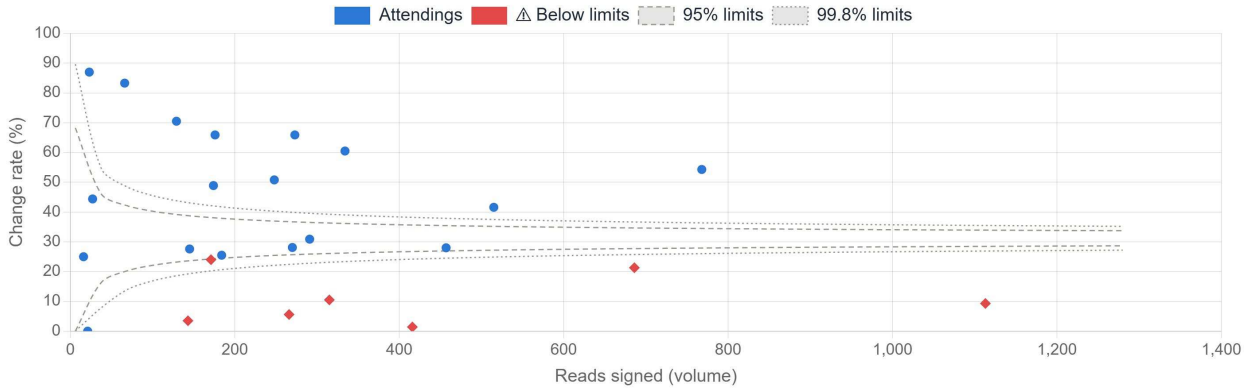

Figure 1: Funnel plot mapping each attending change rate versus how many draft reports CXRs they signed. The funnel curves are the 95% and 99.8% control limits. This demonstrates that 28% (7/25) of the attending were outliers. Assuming residents make errors at similar rates across shifts, these radiologists may be at higher risk for automation bias with draft reporting.

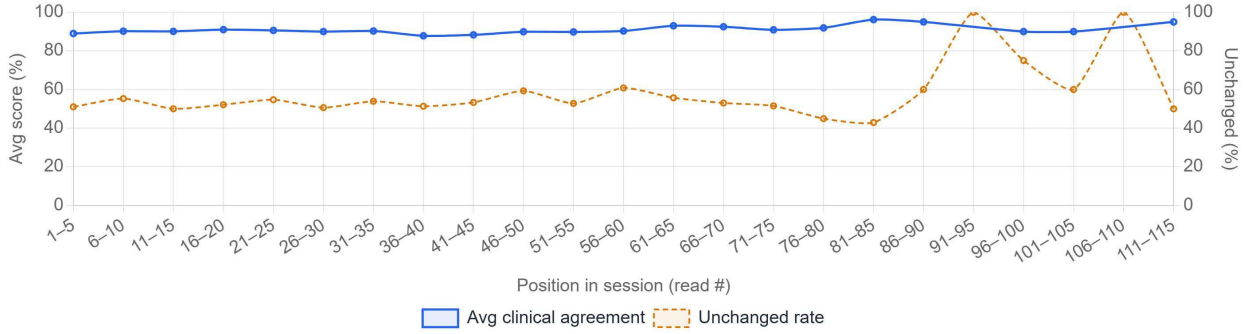

Figure 2: Average LLM scores for clinical agreement and the average percentage unchanged binned by position in shift demonstrating increases in percentage of unchanged reports at higher volumes. This indicates that radiologists are at a higher risk of automation bias at higher volumes or toward the end of shifts.

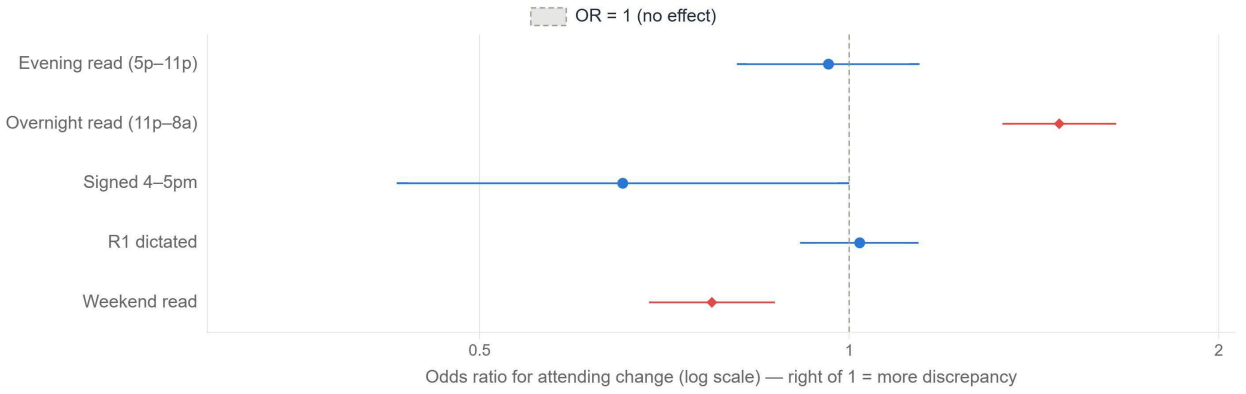

Figure 3: Forrest plots comparing read parameters for report changes highlighting that fewer reports are changed on weekends (OR 0.773, 95% CI 0.687 – 0.870) and more reports are changed overnight (OR 1.483, 95% CI 1.333-1.650). Of note, R1 is not different than R2 dictating within the last two months of R1 year.

Keywords

AI draft reporting; large language model; automation bias