RadGame: An AI-Powered Platform for Radiology Education

Mohammed Baharoon[*,1], Siavash Raissi[*,1], John S. Jun[*,1], Thibault Heintz[*,2,3], Mahmoud Alabbad[4], Ali Alburkani[4], Sung Eun Kim[1,5], Kent Kleinschmidt[6], Abdulrahman O. Alhumaydhi[7], Mohannad Mohammed G. Alghamdi[7], Jeremy Francis Palacio[6], Mohammed Bukhaytan[7], Noah Michael Prudlo[6], Rithvik Akula[6], Brady Chrisler[6], Benjamin Galligos[6], Mohammed O. Almutairi[7], Mazeen Mohammed Alanazi[7], Nasser M. Alrashdi[7], Joel Jihwan Hwang[6], Sri Sai Dinesh Jaliparthi[6], Luke David Nelson[6], Nathaniel Nguyen[6], Sathvik Suryadevara[6], Steven Kim[8], Mohammed F. Mohammed[9], Yevgeniy R. Semenov[10], Kun-Hsing Yu[1], Abdulrhman Aljouie[11,12], Hassan AlOmaish[†,4], Adam Rodman[†,13], Pranav Rajpurkar[†,1]
MOHAMMED_BAHAROON@HMS.HARVARD.EDU

[1] *Department of Biomedical Informatics, Harvard Medical School, Boston, MA*
[2] *Department of Radiation Oncology, Mass General Brigham, Boston, MA*
[3] *Maastricht University, Maastricht, Netherlands*
[4] *Department of Medical Imaging, King Abdulaziz Medical City, Ministry of National Guard, Riyadh, Saudi Arabia*
[5] *National Strategic Technology Research Institute, Seoul National University Hospital, South Korea*
[6] *Saint Louis University School of Medicine, St. Louis, MO*
[7] *College of Medicine, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia*
[8] *Tufts University School of Medicine, Boston, MA*
[9] *Diagnostic Imaging Department, King Faisal Specialist Hospital & Research Center, Riyadh, Saudi Arabia*
[10] *Department of Dermatology, Massachusetts General Hospital, Boston, MA*
[11] *Department of Data Management, King Abdullah International Medical Research Center, Riyadh, Saudi Arabia*
[12] *Department of Health Informatics, King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia*
[13] *Beth Israel Deaconess Medical Center, Boston, MA*

[*] **These authors contributed equally.**
[†] **Senior authors.**

Abstract

We introduce **RadGame**, an AI-powered gamified platform for radiology education that targets two core skills: localizing findings and generating reports. Traditional radiology training is based on passive exposure to cases or active practice with real-time input from supervising radiologists, limiting opportunities for immediate and scalable feedback. RadGame addresses this gap by combining gamification with large-scale public datasets and automated, AI-driven feedback that provides clear, structured guidance to human learners. In *RadGame Localize*, players draw bounding boxes around abnormalities, which are automatically compared to radiologist-drawn annotations from public datasets, and visual explanations are generated by vision-language models for user missed findings. In *RadGame Report*, players compose findings given a chest X-ray, patient age and indication, and receive structured AI feedback based on radiology report generation metrics, highlighting errors and omissions compared to a radiologist's written ground truth report from public datasets, producing a final performance and style score. In a prospective evaluation, participants using RadGame demonstrated a *68%* improvement in localization accuracy compared to *17%* with traditional passive methods and a *31%* improvement in report-writing accuracy compared to *4%* with traditional methods after seeing the same cases. RadGame highlights the potential of AI-driven gamification to deliver scalable, feedback-rich radiology training and reimagines the application of medical AI resources in education.

Keywords: Radiology education, Gamification, Medical AI, Report generation, Localization

Data and Code Availability RadGame is built upon publicly available radiology datasets. For *RadGame Localize*, we used the PadChest-GR dataset (de Castro et al., 2025), which contains chest radiographs with radiologist-annotated bounding boxes. For *RadGame Report*, we used the ReXGradient-160K dataset (Zhang et al., 2025), which provides X-rays paired with radiologist-written reports. Both datasets are freely available to the research community under their respective licenses.

The RadGame platform code is avaliable at https://github.com/siavashraissi/RadGame.

Institutional Review Board (IRB) This study was reviewed by the Harvard Faculty of Medicine Institutional Review Board (Protocol #IRB25-0694) and determined to be exempt under 45 CFR 46.104(d)(2)(3).

1. Introduction

Radiology trainees must acquire two fundamental skills: accurately identifying abnormalities on imaging studies and articulating findings in clear, structured reports. Traditional approaches to radiology education rely on didactic lectures, passive exposure to cases, and supervised readings (Griffith et al., 2019). These methods provide limited opportunities for immediate, personalized feedback. Prior work in medical education has shown that active learning improves diagnostic accuracy and knowledge retention, suggesting the need for more interactive and feedback-rich training methods in radiology (Freeman et al., 2014). Moreover, recent work on artificial intelligence (AI) augmented education further highlights its potential to deliver adaptive, real-time feedback and personalized learning experiences across medical training settings (Shaw et al., 2025; Hui et al., 2025).

However, current radiology education platforms fall short in two key ways. First, most lack real-time, structured, and personalized feedback: trainees can review cases or observe ground-truth annotations, but rarely receive automated guidance tailored to their specific errors or skill level (Duong et al., 2019; Griffith et al., 2019). Second, existing platforms often rely on small, highly curated datasets for simplified tasks that do not capture the diversity, complexity, or volume of real-world radiology interpretation (Biswas et al., 2022; Banerjee et al., 2023; Ali et al., 2021). This limits their ability to provide trainees with the breadth of exposure and adaptive learning experiences needed to prepare for real-world clinical practice.

To address these limitations, we report the following contributions: (1) We develop *RadGame*, an AI-powered gamified platform that teaches the two core radiology tasks—localization and report writing—by repurposing existing large-scale AI radiology datasets and evaluation metrics to provide structured feedback for human learners. (2) We conduct a prospective, multi-institutional user study, showing that using RadGame is associated with improvements in localization and report-writing performance compared to traditional passive learning. (3) We develop *CRIMSON*, an extension of GREEN motivated by RadGame's role as a human-centered evaluation framework. RadGame revealed GREEN's limitations in accounting for clinical context (e.g., age, indication) when measuring the clinical significance of errors, leading to CRIMSON as a more context-aware metric.

2. Related Works

Radiology education has traditionally relied on static textbooks, didactic lectures, and limited opportunities for interactive feedback. In recent years, there has been growing interest in incorporating gamification and personalized training paradigms to better align with the needs of modern trainees and the cognitive demands of radiologic interpretation (Duong et al., 2019; Hui et al., 2025; Biswas et al., 2022).

Several efforts have explored gamified systems for radiology learning. Banerjee et al. introduced RADHunters, a first-person game simulating nodule detection tasks in chest CT imaging, showing that gamification can enhance perceptual skill acquisition and learner engagement (Banerjee et al., 2023). Similarly, SonoGames utilized competitive ultrasound-based quizzes in residency training programs, demonstrating improvements in knowledge retention (Ali et al., 2021). Winkel et al. (2020) evaluated a gamified e-learning platform for pneumothorax detection, where timed challenges with immediate feedback significantly increased diagnostic confidence and skill retention among radiology residents. Additionally, Mobley et al. (2023) reviewed the role of the Kaizen platform, an app-based educational tool with gamified multiple-choice formats, instant feedback, and competitive leaderboards to motivate learners. Early evidence suggests Kaizen enhances knowledge reten-

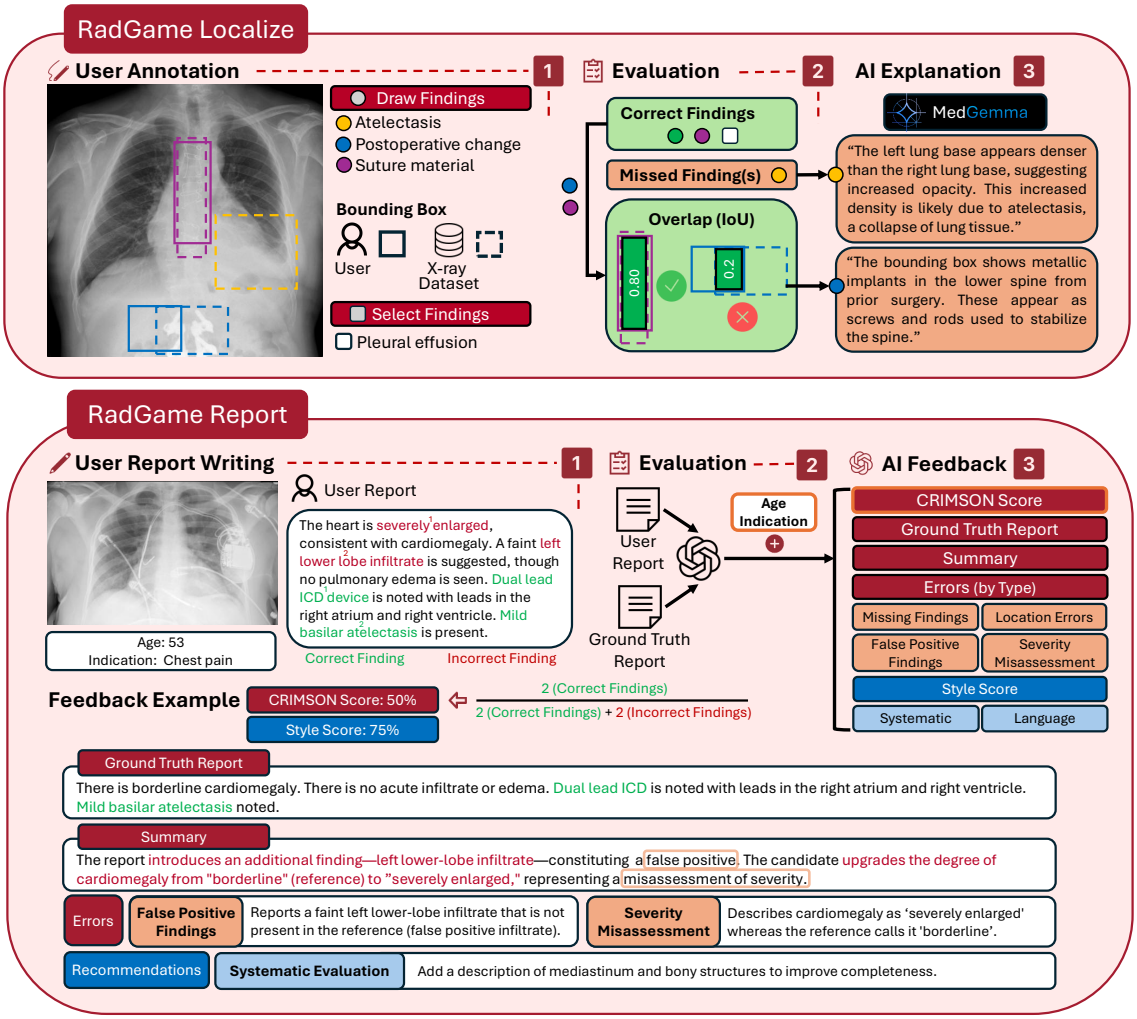


Figure 1: **Overview of RadGame's User Workflow.** In *Localize*, users identify chest X-ray findings either by drawing bounding boxes for location-dependent abnormalities (Draw findings) or by selecting findings that are consistently associated with a fixed anatomical region or that cannot be localized (Select findings). For all existing findings, ground truth bounding boxes are overlaid, and MedGemma 4B generates explanations for findings that are missed or incorrectly identified. A finding is considered correct only if the IoU is over 0.25. In *Report*, users draft finding reports that are assessed by a GPT-o3 using CRIMSON (see Section 5), producing structured outputs that include the CRIMSON score, ground truth findings, summary, and categorized errors. A Style Score is also produced, which covers the report's completeness across all major chest X-ray regions (lungs, heart, bones, mediastinum) and the use of full sentences and clinical language.

tion, peer interaction, and engagement among radiology trainees.

Recent work has also explored integrating artificial intelligence into radiology education (Hui et al., 2025; Wang et al., 2024; Cheng et al., 2020; Saricilar et al., 2023). Biswas et al. (2022) demonstrated the potential of AI-augmented education through a web-based application for chest X-ray nodule detection, showing that AI can deliver real-time, interactive feedback to enhance perceptual training. Similarly, Cheng et al. (2020) showed that an AI-assisted education system significantly improved medical stu-

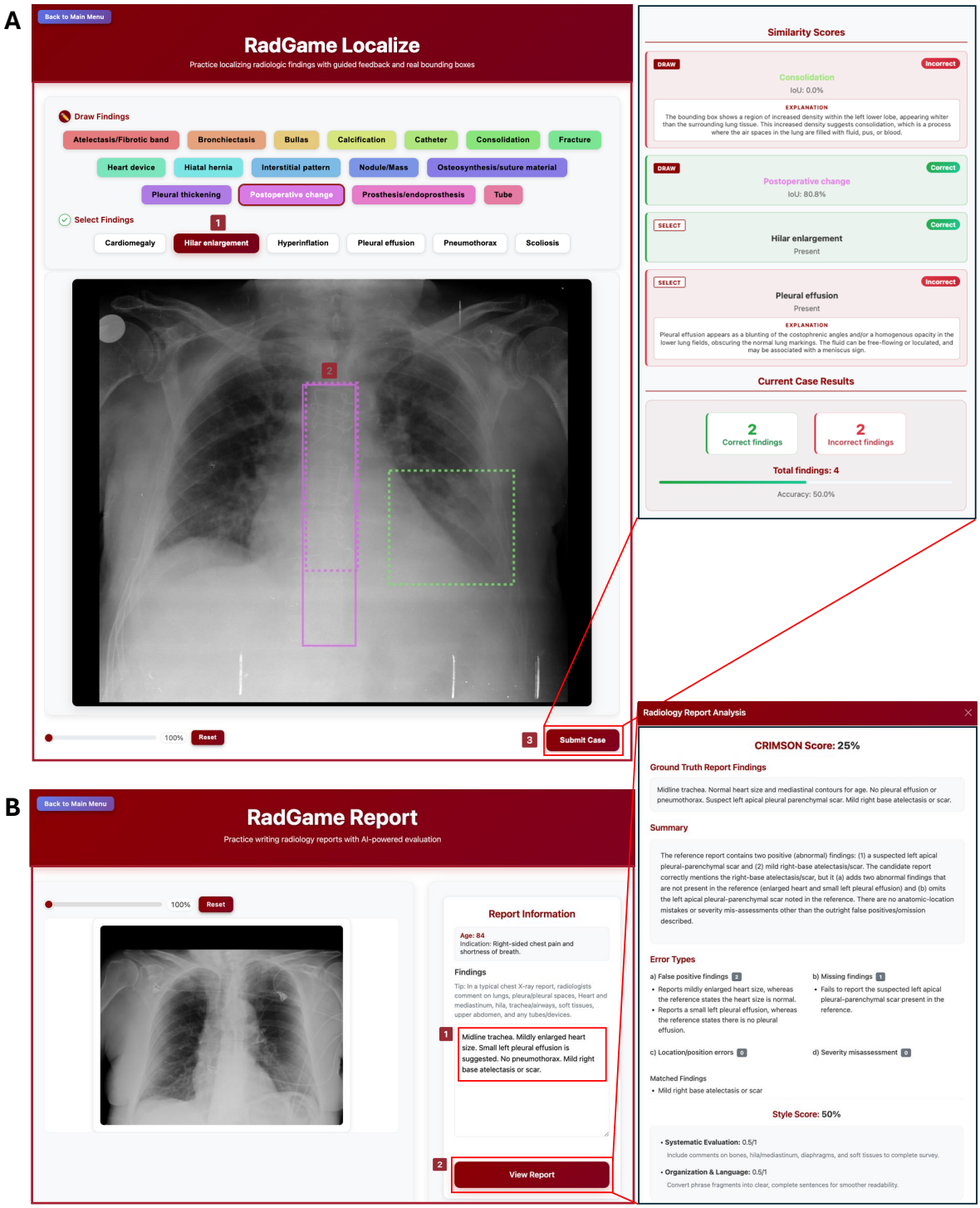


Figure 2: **RadGame User Interface.** Screenshot of the RadGame platform showing both modules: **(A)** *Localize*, where users identify findings on chest X-rays either by drawing bounding boxes or selecting predefined options, and **(B)** *Report*, where users compose findings reports given the image, age, and indication.

dents' diagnostic accuracy for hip fracture detection, particularly among students with lower baseline performance.

Building on these findings, RadGame proposes the first AI-driven educational platform that combines interactive finding localization across 22 finding class types and report writing with automated, personalized feedback at scale.

3. RadGame

RadGame is an AI-powered gamified platform that integrates two core radiology tasks—report generation and finding localization—into interactive learning modules. The platform leverages large-scale public datasets and AI-based feedback mechanisms to provide trainees with structured, personalized evaluation at scale. RadGame consists of two modules: *RadGame Localize* and *RadGame Report*. Figure 1 shows the overall workflow of the two modules.

3.1. RadGame Localize

In *RadGame Localize*, participants are shown single frontal chest X-rays from the PadChest-GR dataset (de Castro et al., 2025) and are prompted to identify different radiological findings. Since finding labels in PadChest-GR come directly from reports, they were originally very specific, with classes like "lobar atelectasis" and "segmental atelectasis," which led to some labels having a very small occurrence count. These findings were combined into a single class (for example, "atelectasis"). They are then divided into two categories: *Draw Findings* and *Select Findings*. Supplementary Figure 2 shows all the classes.

Draw Findings are abnormalities that require the trainee not only to identify their presence but also to localize them by drawing a bounding box on the image. These findings are typically focal and location-dependent, meaning they can appear in different regions of the chest ("Nodule/Mass", "Fracture", "Calcification", etc.). A prediction is considered correct if the intersection-over-union (IoU) between the trainee's bounding box and the radiologist's ground-truth annotation exceeds 0.25, a threshold determined in consultation with radiologists. *Select Findings*, in contrast, are findings where trainees only need to indicate their presence from a checklist without drawing a box. These are typically diffuse abnormalities or findings that are consistently associated with a fixed anatomical location ("Cardiomegaly", "Scoliosis", etc.).

When a trainee misses a finding, visual feedback is provided through MedGemma 4B (Sellergren et al., 2025). For Draw Findings, the system overlays the ground-truth bounding box on the image and generates a two-sentence explanation describing how the missed abnormality can be visually recognized. For Select Findings, MedGemma provides a general description of the typical radiographic appearance of the missed finding. An example is shown in Figure 2A. In Appendix A we provide a validation of MedGemma's explanations.

3.2. RadGame Report

In *RadGame Report*, trainees are presented with all images from a chest X-ray study and prompted to write a radiology report. The ground-truth reference is derived from the *Findings* section of the ReXGradient-160K dataset (Zhang et al., 2025). Studies that have priors or comparisons are excluded. Submitted reports are automatically evaluated using the CRIMSON metric (Section 5), a report generation metric adapted from GREEN (Ostmeier et al., 2024). Unlike GREEN, CRIMSON ignores normal findings that would otherwise inflate scores and incorporates patient age and clinical indication to weigh the clinical significance of errors. Formally, CRIMSON is defined by adapting the GREEN score formula:

$$\text{Score} = \frac{\#\,\text{matched findings}}{\#\,\text{matched findings} \;+\; \sum_{i=(a)}^{(d)} \#\,\text{error}_{\text{sig},i}}$$

where # matched findings is the number of findings that appear in both the candidate and the reference report and errors a through d corresponding to the four different error types (see Figure 1). Errors e (mentioning a comparison that isn't in the reference) and f (omitting a comparison detailing a change from a prior study) from GREEN are dropped since they are concerned with priors.

We use GPT-o3 as our LLM model to generate the report evaluation. The evaluation output includes a single CRIMSON score between 0 and 100%, the ground truth report for comparison, a structured summary of the trainee's submission and errors, and CRIMSON error categories that cover four error categories: false positives, missing findings, location or position errors, and severity misclassification. Figure 2B shows an example of this.

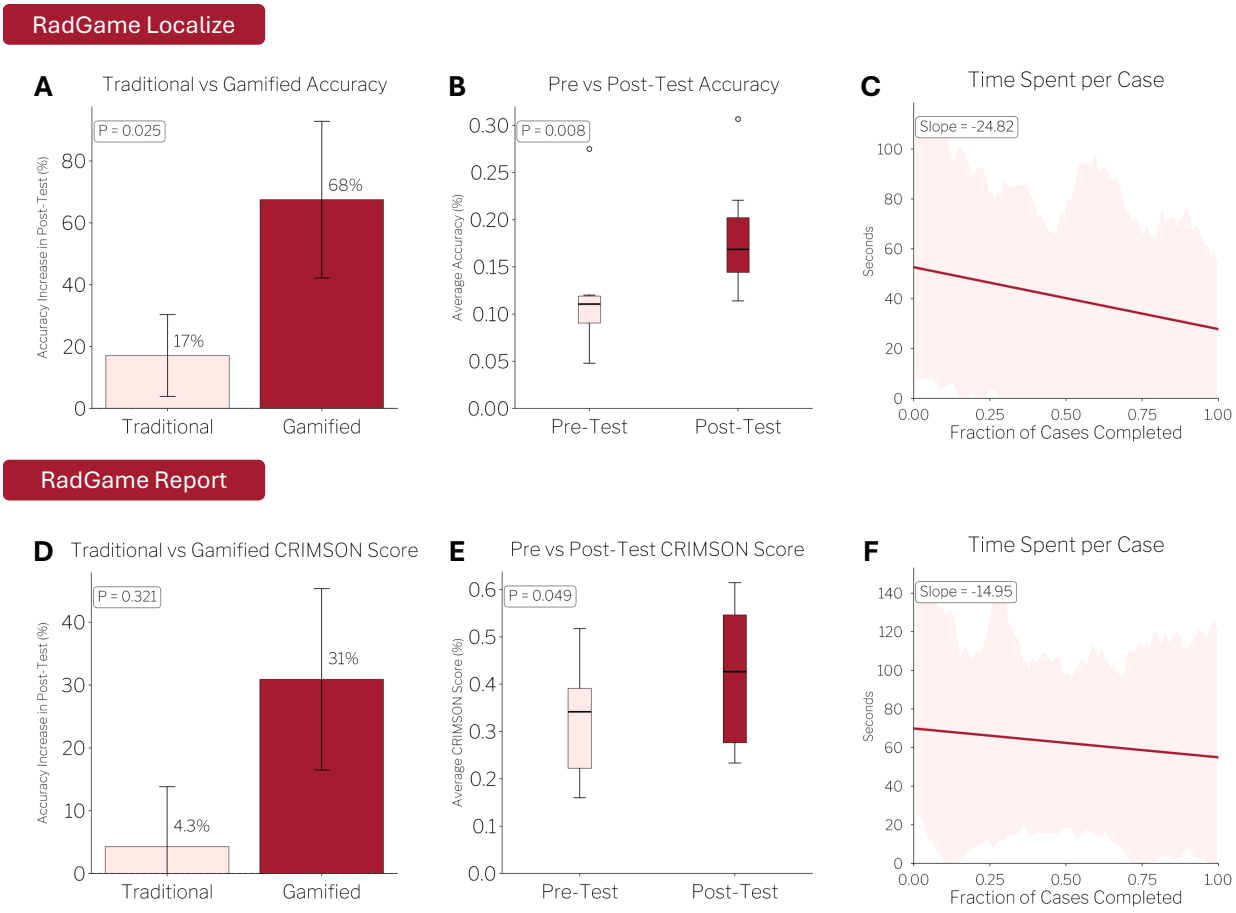


Figure 3: **Performance and efficiency improvements with RadGame across both modules.** The top row shows results for RadGame Localize: **(A)** comparison of accuracy improvements between Gamified and Traditional groups, **(B)** pre-test vs. post-test accuracy changes, and **(C)** reduction in time spent per case over training. The bottom row shows corresponding results for RadGame Report: **(D)** accuracy improvements in Gamified vs. Traditional groups, **(E)** pre-test vs. post-test CRIMSON score changes, and **(F)** reduction in time spent per case. Statistical significance for Gamified vs. Traditional comparisons was assessed using a two-tailed Mann–Whitney U test, while pre- vs. post-test comparisons were assessed using a one-tailed Wilcoxon signed-rank test. Error bars represent the Standard Error of the Mean.

The evaluation output also includes a "Style Score", scaled from 0 (a poorly styled report) to 100% (a professionally styled report), encouraging trainees to use best practices in radiology reporting. The style score considers whether the report covers all major chest X-ray regions (lungs, heart, bones, mediastinum), as well as the organization of the report including use of complete sentences and clinical language (see Supplementary Figure 7 for the prompt).

3.3. Customizability

RadGame is designed with flexibility in mind, allowing adaptation across multiple dimensions. The platform can be extended to incorporate different datasets, additional finding categories, and alternative feedback strategies. For the localization task in particular, RadGame supports customizing the set of *Draw* and *Select* findings to match the characteristics of the dataset or the learning objectives. For exam-

ple, the platform can be easily extended to support a version for localizing nodules only.

To further support this extensibility, we developed a specialized version of RadGame focused on interstitial lung patterns, which are subtle and diagnostically challenging findings. For this module, we created the first bounding-box annotated dataset based on ReXGradient-160k (Zhang et al., 2025) specifically aimed at differentiating between distinct interstitial patterns, including Kerley lines, miliary, and reticulonodular opacities. The distribution of findings for this dataset is shown in Supplementary Table 3. The dataset will be made publicly available.

4. User Study

4.1. Study Design

Cohort. Eighteen medical students were selected to participate in the user study (see Supplementary Table 1 for demographics). Participants come from Saint Louis University (Saint Louis, MO), Tufts University School of Medicine (Boston, MA), and King Saud bin Abdulaziz University for Health Sciences (Riyadh, Saudi Arabia), comprising a multinational, multi-institutional cohort. Students were selected based on their interest in radiology. At the time of participation, nine students were in the pre-clinical stage (M1, M2) of medical training and nine were in the clinical stage (M3, M4). When asked about prior radiology experience across classes and clinical rotations, seven students (38.8%) reported no experience, five (27.8%) had only taken radiology classes, two (11.1%) had only performed radiology rotations, and four (22.2%) had experience with both.

Group Assignment. Participants first completed the *RadGame Localize* module, followed by *RadGame Report* module. They were randomly assigned to either a *Gamified* or *Traditional* feedback group for each module, with a crossover design such that participants assigned to the Gamified group for Localize were assigned to the Traditional group for Report, and vice versa. This resulted in 8 and 10 participants completing Localize and Report, respectively, in the Gamified group.

Participants assigned to the Gamified group received AI-generated explanations and context-aware error feedback, while those in the Traditional modality learned by observing cases and ground truths without interactive feedback (see Supplementary Figure 5). For *RadGame Localize*, the Gamified group drew bounding boxes with AI-generated explanations for user missed findings (see Figure 2A), whereas the Traditional group viewed only the case with ground-truth boxes overlaid. For *RadGame Report*, the Gamified group wrote reports with full automated and personalized feedback, while the Traditional group viewed only the chest X-ray and the associated ground-truth report (see Figure 2B). Both groups saw the same cases in the same order.

Study phases. The study was conducted in three phases: a baseline pre-test, a learning phase with the assigned group, and a post-test with identical cases to the pre-test for final evaluation. Participants were asked to complete the three components within a seven-day period and did not receive their test scores during the pre- and post-tests.

In RadGame Localize, the pre-test and post-test consisted of completing 25 RadGame Localize cases completed within a 45-minute timeframe. A senior radiologist selected the test cases from PadChest-GR to evenly distribute case difficulties (de Castro et al., 2025). To ensure that the ground truth annotations of the test cases were as accurate and consistent as possible, two senior radiologists were assigned to re-evaluate the findings and their respective annotations—revising bounding boxes while adding missed conditions. Upon completion of the pre-test, participants were assigned to complete 375 cases of RadGame Localize in their assigned group. Once participants completed all cases, they were asked to take the post-test and move to RadGame Report.

In RadGame Report, the pre-test and post-test consisted of completing 10 RadGame Report cases within a 45-minute timeframe. A senior radiologist selected the test cases from ReXGradient-160K to establish an even distribution of case difficulties (Zhang et al., 2025). Ground-truth and student reports for both tests were reviewed by radiologists to ensure that final scores reliably measured students' abilities to write reports, rather than optimize CRIMSON scores. The review procedure is described further in Appendix B. Upon completion of the pre-test, participants were assigned to complete 150 cases of RadGame Report in their assigned group. Once participants completed all cases, they were asked to take the post-test to complete the study.

4.2. Results

We evaluated the impact of RadGame on diagnostic accuracy and efficiency across both the *Localize* and

Report modules (Figure 3). Across both modules, participants assigned to the *Gamified* group demonstrated larger performance gains than those in the *Traditional* modality.

For RadGame Localize, participants in the Gamified group showed a 68% improvement in post-test accuracy relative to baseline, compared to only 17% in the Traditional group ($p < 0.05$, two-tailed Mann–Whitney U). Figure 3B shows a comparison between pre- vs. post-test gains for the Gamified group ($p < 0.05$, one-tailed Wilcoxon signed-rank). Specifically, the pre-test showed a median accuracy of 0.111 (range: 0.048–0.275, 95% CI: 0.076–0.120), while the post-test showed a median of 0.169 (range: 0.114–0.307, 95% CI: 0.138–0.221), corresponding to a Cliff's Delta of 0.673 compared to the pre-test. These results demonstrate that interactive bounding box annotation with AI-generated feedback was associated with improved localization skills compared to traditional passive review of ground-truth annotations. In Supplementary Figure 1, we also show localize accuracies across different IoU thresholds, and Supplementary Figure 4 shows accuracy differences in the pre- and post-test across finding labels.

For RadGame Report, the Gamified group showed a 31% improvement in CRIMSON scores from pre- to post-test, compared to 4.3% in the Traditional group. While this between-group difference did not reach statistical significance, this is likely due to the small sample size (n = 18) despite the large observed difference. Figure 3E revealed significant pre- vs. post-test improvements in the Gamified group ($p < 0.05$). Specifically, the pre-test showed a median CRIMSON score of 0.342 (range: 0.160–0.518, 95% CI: 0.222–0.412), while the post-test showed a median of 0.426 (range: 0.233–0.615, 95% CI: 0.268–0.547), with a Cliff's Delta of 0.383 compared to the pre-test. These findings demonstrate that receiving personalized feedback from RadGame may be associated with improved report writing skills.

Finally, across both modules, participants in the Gamified group demonstrated progressive reductions in time spent per case as training advanced, indicating improved diagnostic efficiency alongside accuracy gains (Figure 3C and F). At the end of the training, participants were around 25 seconds faster per case at RadGame Localize and 15 seconds faster per case at RadGame Report. A plot comparing time spent for Gamified and Traditional groups is shown in Supplementary Figure 3.

5. RadGame as a Humanistic Evaluation of AI Material

Beyond its role as an educational tool, *RadGame* can act as a human-in-the-loop evaluation harness for AI radiology material—analogous to Chatbot Arena for large language models (Chiang et al., 2024). In our setting, trainees interact directly with material used to train and evaluate AI models and provide implicit and explicit judgments through gameplay, which serves as a humanistic complement to correlation-based benchmarks.

This was exemplified during our user study, where user feedback post-study revealed two key shortcomings of the GREEN metric, which was developed to evaluate AI report generation capabilities (Ostmeier et al., 2024). First, GREEN rewarded the reporting of normal findings, inflating scores even when significant abnormalities were missed. However, the inclusion of specific normal findings in a report is often determined by the individual radiologist's style rather than by clinical necessity, making such matches a poor basis for evaluation of clinical accuracy. Instead, we believe that it makes more sense to keep a separate score for evaluating report style. This led to us proposing "Style Score," a separate scoring system that evaluates a report along two pillars: (1) Systematic Evaluation and (2) Organization and Language. Systematic evaluation verifies if the report systematically evaluates all important regions in a chest x-ray, including the lungs, heart, bones, and mediastinum. Organization and Language evaluates the report for the use of clinical language, writing in full sentences, and having a generally organized report. The prompt can be found in Supplementary Figure 7. Second, GREEN fails to account for clinical context, penalizing omissions of irrelevant findings (e.g., age-related degenerative changes) while giving equal weight to errors that were clinically consequential.

Guided by these insights, we introduced CRIMSON, a novel metric that (1) ignores matches on normal findings and (2) incorporates age and indication to assess the clinical significance of errors. Figure 4 illustrates these improvements. The prompt for CRIMSON can be found in Supplementary Figure 6. In this way, RadGame not only trains human learners but also functions as a testbed for refining AI evaluation frameworks, bridging the gap between quantitative evaluations and human judgment.

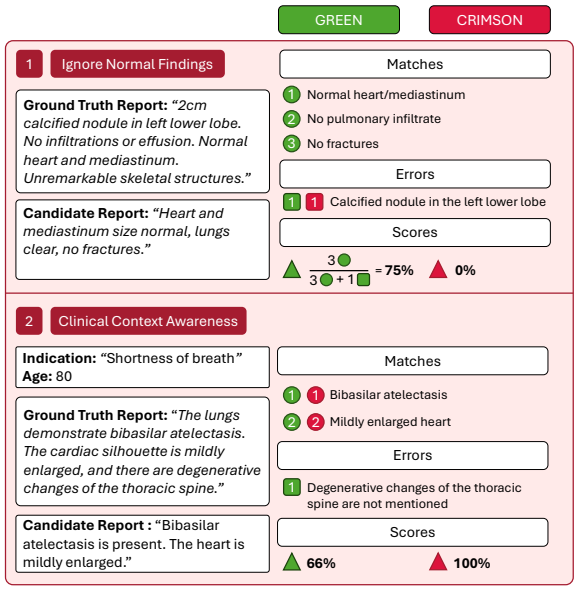


Figure 4: **Comparison of GREEN and CRIMSON scoring.** (1) *Ignore Normal Findings:* GREEN rewards normal findings (e.g., normal heart/mediastinum, no infiltrate, no fractures), inflating the score despite missing the clinically important calcified nodule. CRIMSON excludes such credit, yielding 0%. (2) *Clinical Context Awareness:* For an 80-year-old with shortness of breath, GREEN penalizes omission of degenerative spine changes, whereas CRIMSON deems them insignificant, giving full credit for bibasilar atelectasis and mild cardiomegaly.

6. Discussion

This work introduces RadGame, an AI-powered, gamified platform for radiology education that integrates real-time, structured feedback into both localization and report-writing tasks. Our findings demonstrate that gamified, feedback-rich training is associated with significantly larger gains in diagnostic accuracy and efficiency compared to traditional passive learning methods. These results align with a growing body of literature showing that active learning paradigms and immediate feedback may help with diagnostic reasoning and skill retention across medical education domains (Freeman et al., 2014; Duong et al., 2019; Banerjee et al., 2023). By repurposing large-scale, publicly available radiology datasets and AI evaluation resources into human-centered training experiences, RadGame builds upon earlier efforts in gamification and AI-augmented education while uniquely scaling structured feedback to diverse tasks.

Future efforts will expand RadGame to incorporate additional imaging modalities and tasks beyond chest X-rays. In particular, grounding datasets such as ReXGroundingCT (Baharoon et al., 2025) offer an opportunity to develop RadGame modules for volumetric CT imaging. Such expansions could support training in cross-sectional imaging and enable assessment of three-dimensional spatial reasoning, which is inherently more difficult. Another direction involves making the platform more interactive through back-and-forth dialogue between the trainee and the AI system. Rather than providing static feedback after a single submission, future iterations could allow learners to ask clarifying questions, receive hints in real time, and iteratively refine their reports or localizations based on the AI's guidance. Such a conversational framework would transform RadGame into a more adaptive, tutor-like experience, fostering deeper reasoning and personalized learning trajectories.

This study has several limitations. First, the sample size was relatively small (n = 18), which may limit statistical power for some comparisons. However, we partly mitigate this by including a multi-institutional cohort with participants at different stages of medical training and diverse prior experiences in radiology, improving the representativeness of our findings. Second, we observed that participants in the Traditional group progressed through the cases more quickly than those in the Gamified group. This difference reflects the passive nature of the Traditional modality, where participants simply observed the cases and ground-truth annotations without the additional interactive steps required in the Gamified setting.

Acknowledgments

This project started as a group project from the BMI 702: Foundations of Biomedical Informatics II course at Harvard University. We thank Dr. Marinka Zitnik for her guidance and support throughout the project and course.

References

Maria Fatima Ali, Naila Nadeem, Farah Khalid, Naveed Muhammad Anwar, Ghulam Nabie, and Charles Docherty. Sonogames: sounds of the right kind introducing gamification into radiology training. *BMC Research Notes*, 14(1):341, 2021.

Mohammed Baharoon, Luyang Luo, Michael Moritz, Abhinav Kumar, Sung Eun Kim, Xiaoman Zhang, Miao Zhu, Mahmoud Hussain Alabbad, Maha Sbayel Alhazmi, Neel P Mistry, et al. Rexgroundingct: A 3d chest ct dataset for segmentation of findings from free-text reports. *arXiv preprint arXiv:2507.22030*, 2025.

Soham Banerjee, Rishabh Agarwal, and William F Auffermann. Radhunters: gamification in radiology perceptual education. *Journal of Medical Imaging*, 10(S1):S11905–S11905, 2023.

Som Subhro Biswas, Srirupa Biswas, Sandeep Singh Awal, and Hitesh Goyal. Current status of radiology education online: a comprehensive update. *SN comprehensive clinical medicine*, 4(1):182, 2022.

Chi-Tung Cheng, Chih-Chi Chen, Chih-Yuan Fu, Chung-Hsien Chaou, Yu-Tung Wu, Chih-Po Hsu, Chih-Chen Chang, I-Fang Chung, Chi-Hsun Hsieh, Ming-Ju Hsieh, et al. Artificial intelligence-based education assists medical students' interpretation of hip fracture. *Insights into Imaging*, 11(1):119, 2020.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I Jordan, Joseph E Gonzalez, et al. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, pages 8359–8388, 2024.

Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7): AIdbp2401120, 2025.

Michael Tran Duong, Andreas M Rauschecker, Jeffrey D Rudie, Po-Hao Chen, Tessa S Cook, R Nick Bryan, and Suyash Mohan. Artificial intelligence for precision education in radiology. *The British journal of radiology*, 92(1103):20190389, 2019.

Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences*, 111(23):8410–8415, 2014.

Brent Griffith, Nadja Kadom, and Christopher M Straus. Radiology education in the 21st century: threats and opportunities. *Journal of the American College of Radiology*, 16(10):1482–1487, 2019.

Muying Lucy Hui, Ethan Sacoransky, Andrew Chung, and Benjamin YM Kwan. Exploring the integration of artificial intelligence in radiology education: A scoping review. *Current Problems in Diagnostic Radiology*, 54(3):332–338, 2025.

Alisa Mobley, Agni Chandora, and Stefanie Woodard. The impact of gamification and potential of kaizen in radiology education. *Clinical Imaging*, 103:109990, 2023.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Md, Michael Moseley, Curtis Langlotz, Akshay Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, 2024.

Erin C Saricilar, Annette Burgess, and Anthony Freeman. A pilot study of the use of artificial intelligence with high-fidelity simulations in assessing endovascular procedural competence independent of a human examiner. *ANZ Journal of Surgery*, 93 (6):1525–1531, 2023.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Kody Shaw, Marcus A Henning, and Craig S Webster. Artificial intelligence in medical education: a scoping review of the evidence for efficacy and future directions. *Medical Science Educator*, pages 1–14, 2025.

DongXu Wang, BingCheng Huai, Xing Ma, BaiMing Jin, YuGuang Wang, MengYu Chen, JunZhi Sang, and RuiNan Liu. Application of artificial intelligence-assisted image diagnosis software based on volume data reconstruction technique in medical imaging practice teaching. *BMC Medical Education*, 24(1):405, 2024.

David J Winkel, Philipp Brantner, Jonas Lutz, Safak Korkut, Sebastian Linxen, and Tobias J Heye. Gamification of electronic learning in radiology education to improve diagnostic confidence and reduce error rates. *American Journal of Roentgenology*, 214(3):618–623, 2020.

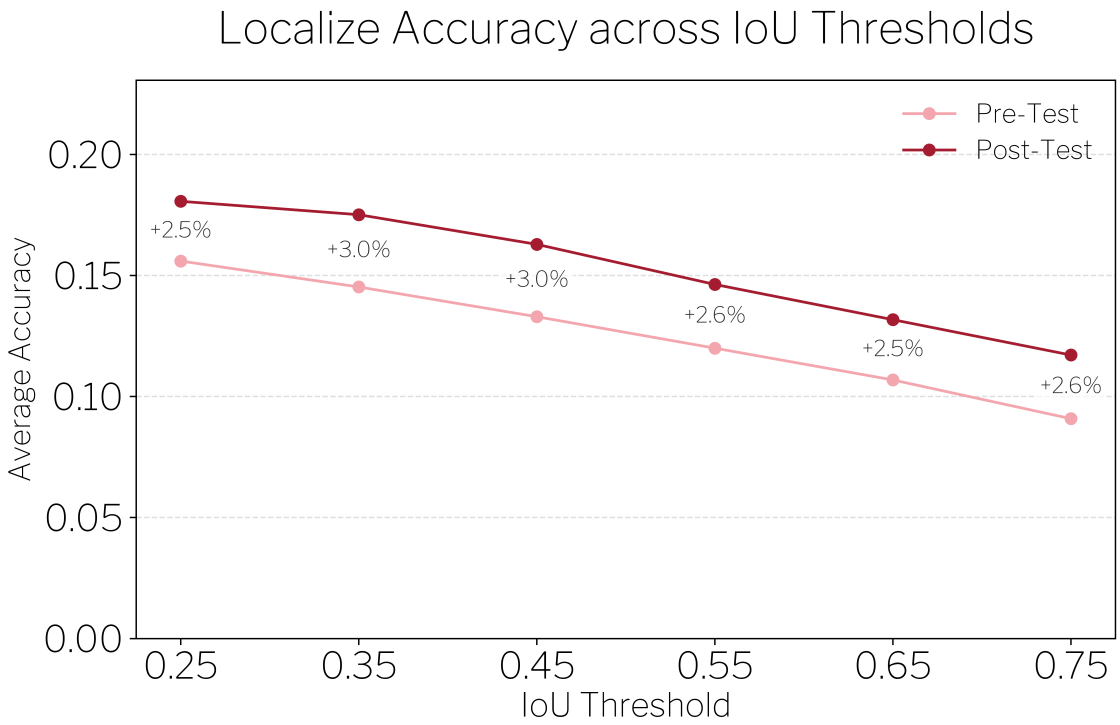
Xiaoman Zhang, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. *arXiv preprint arXiv:2505.00228*, 2025.

Appendix A. MedGemma 4B Explanations Validation

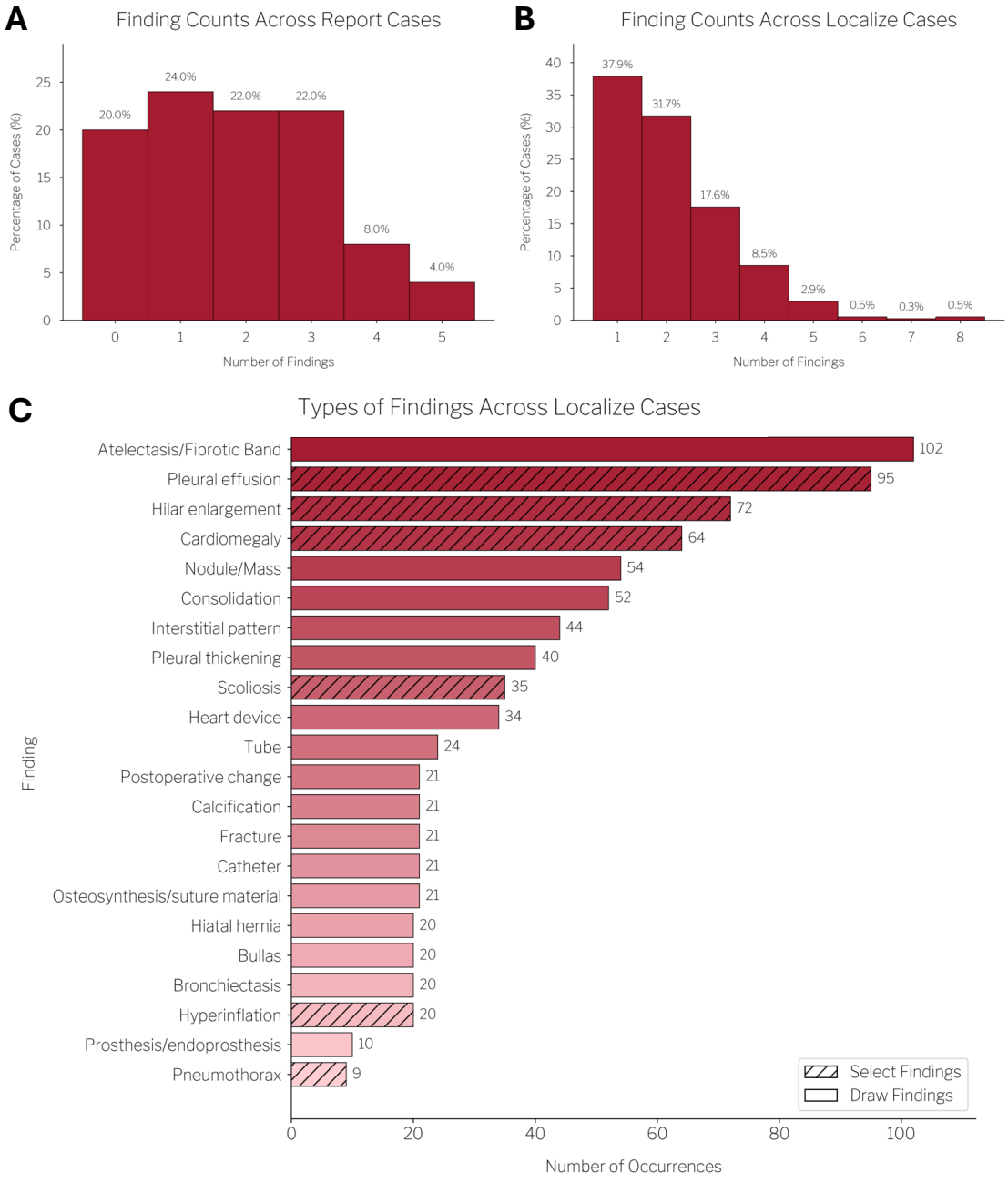
To validate MedGemma's feedback, we sampled 100 random RadGame Localize findings. Two board-certified radiologists reviewed each case and scored MedGemma's descriptions on a binary scale for whether it (1) discussed the correct anatomical location, (2) accurately described the image's visual features, and (3) correctly identified the specific subtype of the finding (e.g., correctly labeling "Pacemaker" when the class is "Heart Device"). MedGemma addressed the correct anatomical location in 88% of findings, correctly identified visual features in 96%, and specified the correct condition present in 90%. We also show performance for this task using two other models, Qwen3VL and Llama3.2, which is shown in Supplementary Table 4. Supplementary Figure 8 shows three examples of such explanations while Supplementary Figure 9 shows three examples of failure cases. Supplementary Figure 10 shows the accuracy of these explanations across finding categories.

Appendix B. RadGame Report Test Review

To ensure that reports in both tests were accurate and did not reference prior studies that would otherwise be unavailable to the participants, two senior board-certified radiologists reviewed and revised all 10 ground-truth reports for comparison. Then, all pre-test and post-test reports written by participants were scored by CRIMSON, followed by manual review by two board-certified radiologists to generate a final score. Radiologists concluded that 3.93% of the total errors were clinically insignificant or could not be determined from the image and were removed, 3.26% reflected correct findings that were penalized by CRIMSON (e.g., labeling an interstitial pattern instead of a Kerley B line), while 0.38% were errors incorrectly deemed insignificant by the model.



Supplementary Figure 1: **Localize Accuracy Across IoU Thresholds.** Post-test accuracies remain consistently higher than pre-test across IoU thresholds.



Supplementary Figure 2: **Distribution of Findings and Finding Counts Across RadGame Localize and Report. (A)** A distribution of the number of findings present across all 150 RadGame Report cases. Cases with a diverse number of findings were selected to adjust the platform's difficulty. **(B)** A distribution of the number of findings present across all 375 RadGame Localize cases. **(C)** Counts of the different types of findings across all 375 RadGame Localize cases, divided into findings that only require a binary selection ("Select Findings") and findings that require a selection and a bounding box ("Draw Findings").

Supplementary Table 1: **Study Cohort Demographics (Counts and Percentages)**

Question	Response	n	%
What is the name of your university/primary affiliation?	St. Louis University	11	61.1
	King Saud bin Abdulaziz University for Health Sciences	6	33.3
	Tufts University School of Medicine	1	5.6
What stage of medical training are you currently in?	Pre-Clinical	9	50
	Clinical	9	50
Have you done a radiology rotation/classes?	Classes	5	27.8
	Rotation	2	11.1
	Both	4	22.2
	None	7	38.8
Determining comfort in radiology - How comfortable are you with interpreting X-ray images on a scale of 1-5? (1 being the least comfortable)	1	0	0
	2	11	61.1
	3	4	22.2
	4	3	16.7
	5	0	0
Determining comfort in radiology - How comfortable are you with PA/AP chest X-rays specifically, on a scale of 1-5? (1 being the least comfortable)	1	2	11.1
	2	4	22.2
	3	7	38.9
	4	3	16.7
	5	0	0

Supplementary Table 2: **RadGame Localize: Draw and Select Findings.**

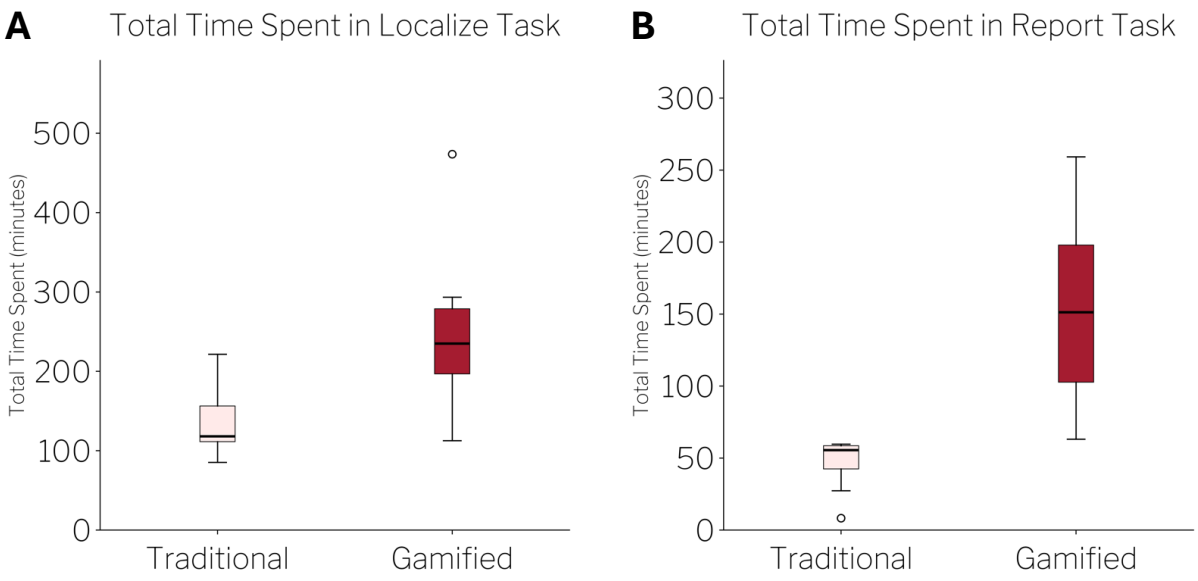
Draw Findings		
1. Atelectasis/Fibrotic band	2. Bronchiectasis	3. Bullas
4. Calcification	5. Catheter	6. Consolidation
7. Fracture	8. Heart device	9. Hiatal hernia
10. Interstitial pattern	11. Nodule/Mass	12. Osteosynthesis/suture material
13. Pleural thickening	14. Postoperative change	15. Prosthesis/endoprosthesis
16. Tube		
Select Findings		
17. Cardiomegaly	18. Hilar enlargement	19. Hyperinflation
20. Pleural effusion	21. Pneumothorax	22. Scoliosis

Supplementary Table 3: **Distribution of Cases Across Interstitial Pattern Subtypes.**

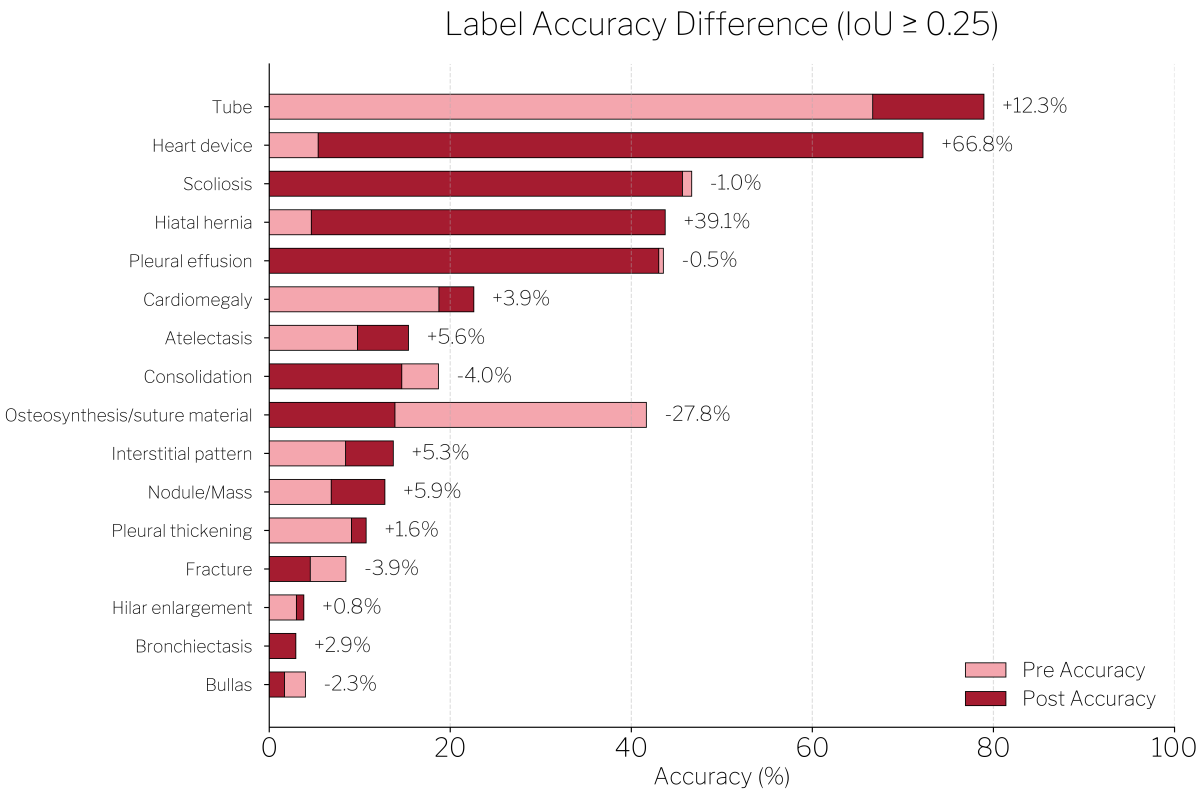
Interstitial Pattern Subtype	**Number of Cases**
Nodular/Miliary	51
Reticulonodular	133
Reticular/Kerley B line	260

Supplementary Table 4: **Comparison of Sample Accuracies Across VLMs.** We use a sample size of 100.

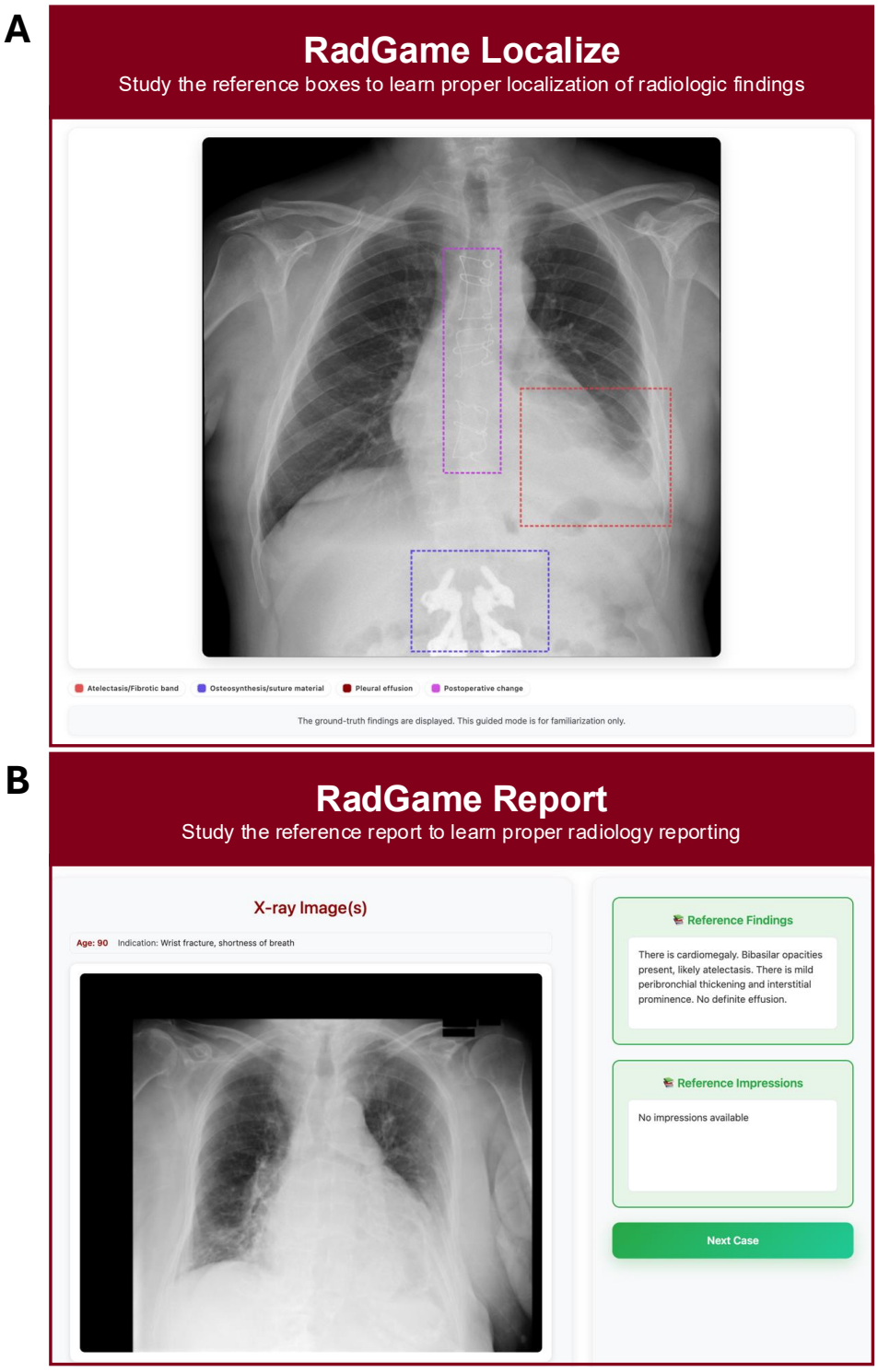
Model	**Location Accuracy**	**Visual Accuracy**	**Class Accuracy**
MedGemma 4B	**88**	**96**	90
Qwen3VL 4B	80	92	94
Llama3.2 11B Vision	81	86	**96**



Supplementary Figure 3: **Localize Accuracy Across Finding Labels.**



Supplementary Figure 4: **Localize Accuracy Difference Pre- and Post-Test Across Finding Labels.**



Supplementary Figure 5: **RadGame User Interface (Traditional Module).** Screenshot of the RadGame platform showing modules in traditional learning mode: **(A)** *Localize*, where users can see findings present in chest X-rays and their respective bounding boxes, and **(B)** *Report*, where users are provided with findings from the ground truth report.

CRIMSON Prompt

Objective: Evaluate the accuracy of a candidate radiology report in comparison to a reference radiology report composed by expert radiologists. Only include positive findings, not normal findings. Do not include notes unrelated to clinical findings.

Process Overview:
You will be presented with:

1. The criteria for making a judgment.
2. The reference radiology report.
3. The candidate radiology report.
4. The desired format for your assessment.

1. Criteria for Judgment: For each candidate report, determine only the clinically significant errors.
   Errors can fall into one of these categories:
   a) False report of a finding in the candidate.
   b) Missing a finding present in the reference.
   c) Misidentification of a finding's anatomic location/position.
   d) Misassessment of the severity of a finding.

   Note: Concentrate on the clinical findings rather than the report's writing style. Evaluate only the findings that appear in both reports.

   Patient Context:
   Age: {age}
   Indication: {indication}

   IMPORTANT NOTES:

   - Evaluate only positive findings, not normal findings. If a finding is normal, it should not be counted in the errors.
   - Ignore all references to prior findings and studies. DO NOT COUNT THEM AS ERRORS.
   - Do NOT penalize the candidate report for omitting specific numeric measurements (e.g., size or dimensions of a nodule/lesion) if the underlying finding is correctly identified. Missing measurements alone is fine since the user writing the candidate report can't measure. They should only be penalized for missing the finding itself.
   - Do NOT penalize omission of age-appropriate findings that are NOT clinically significant in the context of the indication and patient age. For example, if the patient is over 65 years old, do not penalize omission of expected degenerative changes such as aortic calcification, vascular tortuosity, degenerative spine changes, etc, UNLESS it is related to the indication.
   - Do NOT hallucinate or infer findings absent from both reports.
2. Reference Report: {reference}
3. Candidate Report: {candidate}
4. Reporting Your Assessment: Format your output as a JSON. Follow this specific format for your output, even if no errors are found:

```
{
    "Explanation": "<Explanation>",
    "ClinicallySignificantErrors": {
        "a": ["<Error 1>", "<Error 2>", "...", "<Error n>"],
        "b": ["<Error 1>", "<Error 2>", "...", "<Error n>"],
        "c": ["<Error 1>", "<Error 2>", "...", "<Error n>"],
        "d": ["<Error 1>", "<Error 2>", "...", "<Error n>"]},
    "MatchedFindings": ["<Finding 1>", "<Finding 2>", "...", "<Finding n>"]
}
```

Supplementary Figure 6: **The prompt used to generate CRIMSON scores.**

Style Score Prompt

Objective: Evaluate the writing style and structure of a radiology report to determine how well it follows professional radiology reporting standards. Focus on style, structure, and systematic evaluation rather than clinical accuracy.
Criteria for Judgment:

Rate each aspect as 0 (poor), 0.5 (adequate), or 1 (excellent):

1. SYSTEMATIC EVALUATION: Does the report cover the major chest X-ray regions?
   - 1.0: Covers most/all major areas (lungs, heart, bones, mediastinum) in organized way
   - 0.5: Covers several major areas but may miss 1-2 or lack organization
   - 0.0: Only mentions 1-2 areas or very disorganized
2. ORGANIZATION AND LANGUAGE: Is the report reasonably well-organized and written in appropriate clinical language?
   - 1.0: Clear organization with, complete sentences and clinical language
   - 0.5: Some organization present, mostly complete sentences
   - 0.0: Poor organization, incomplete sentences, non-clinical language

Candidate Report: {candidate}

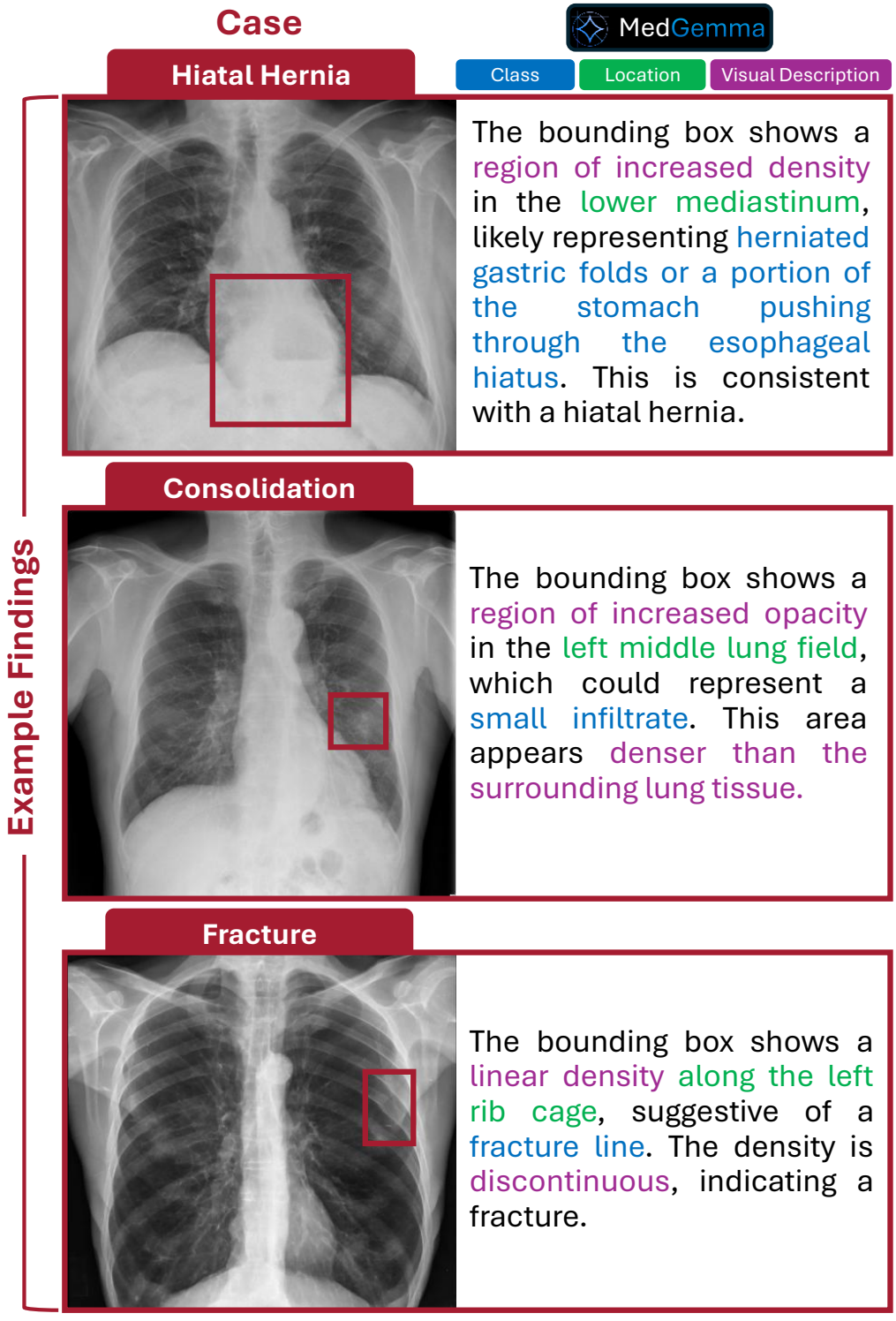
NOTES:

- Do NOT recommend the user to make sections or sub-sections in the report such as Findings, Impression, etc.
- Provide 1 recommendation per scoring category that scored less than 1.0
- If a category scores 1.0 (perfect), leave that recommendation field empty (“”)
- Keep each recommendation very concise and actionable

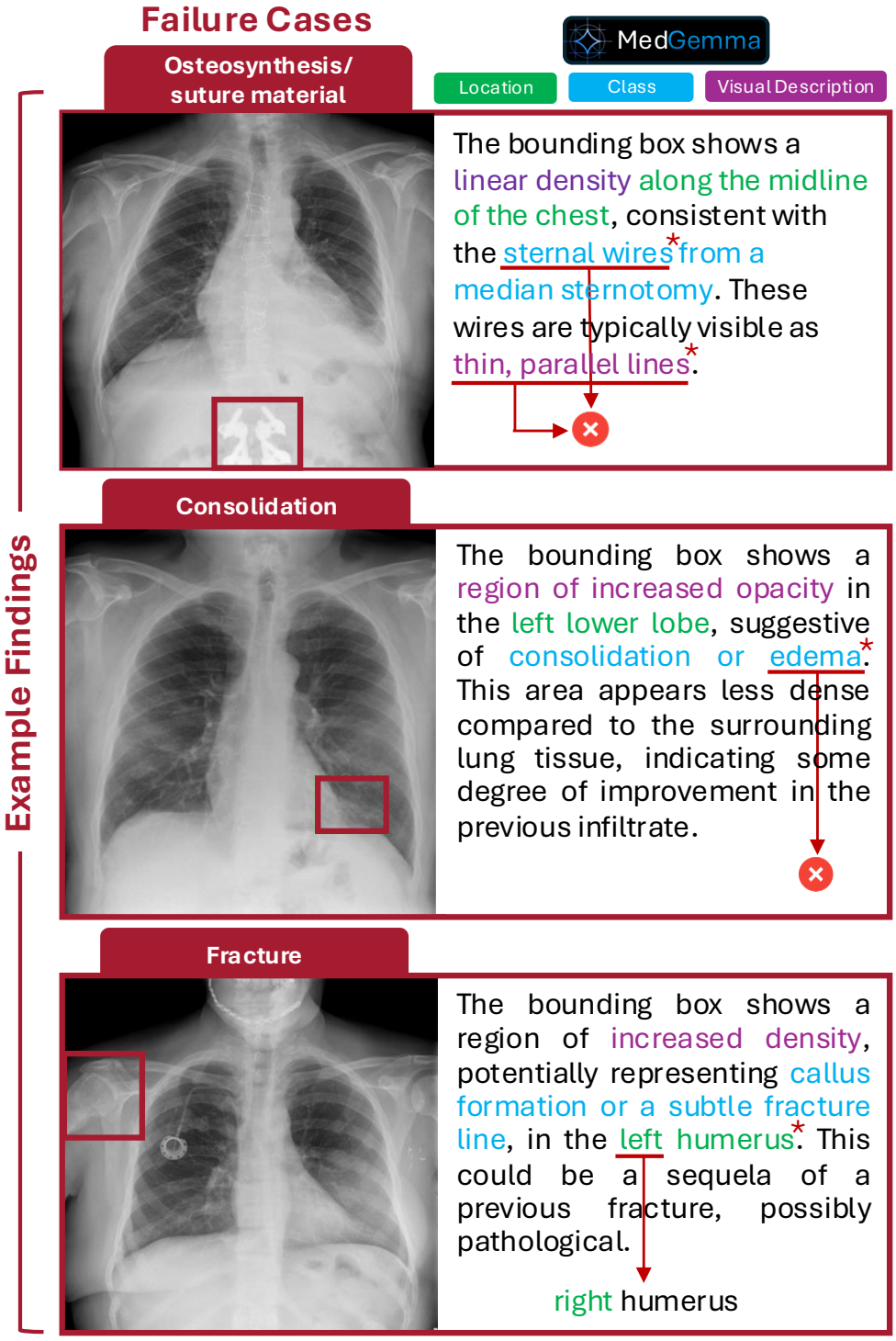
Be concise in your recommendations. Provide your assessment in the following JSON format:

```
{
    “systematic_evaluation_score”: <0, 0.5, or 1>,
    “organization_language_score”: <0, 0.5, or 1>,
    “systematic_evaluation_recommendation”:
        “<Recommendation if score < 1, otherwise empty>”,
    “organization_language_recommendation”:
        “<Recommendation if score < 1, otherwise empty>”
}
```

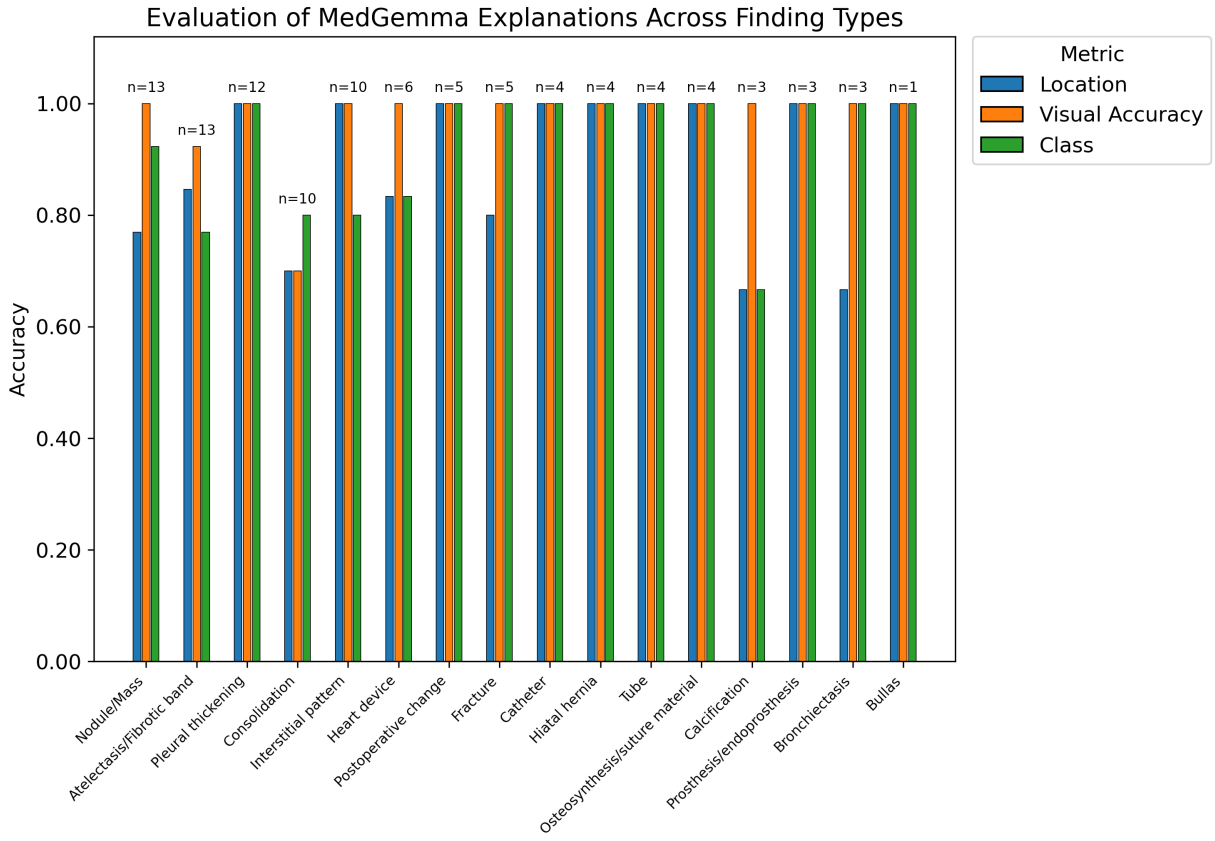
Supplementary Figure 7: **The prompt used by Style Score to evaluate the candidate report on “Systematic Evaluation” and “Organization and Language”.**



Supplementary Figure 8: **Examples of MedGemma Explanations.**



Supplementary Figure 9: **Examples of MedGemma Failure Explanations.**



Supplementary Figure 10: **Accuracy of MedGemma Explanations Across Finding Categories.** The reported accuracy is from 100 cases sampled randomly and evaluated by a senior radiologist.