Journal of Imaging Informatics in Medicine
https://doi.org/10.1007/s10278-026-01960-w

League of Radiologists—an End-to-End AI Framework for Scalable and Gamified Radiology Education: A Pilot Implementation in Chest Radiography

Hyunji Kim[1] · Young-Tak Kim[1] · Saul Langarica[2] · Kevin P. Fialkowski[1] · Jarrel C. Y. Seah[1] · Jennifer S. N. Tang[3] · Kyoung Doo Song[4] · Dae Chul Jung[5] · Kyongtae Tyler Bae[6] · Rory L. Cochran[1] · Marc D. Succi[1,7,8] · Shaunagh McDermott[1] · Manisha Bahl[1] · Jeanne B. Ackman[1] · Michael H. Lev[1] · Michael S. Gee[1] · Synho Do[1,9]

Received: 5 February 2026 / Revised: 11 March 2026 / Accepted: 2 April 2026
© The Author(s) under exclusive licence to Society for Imaging Informatics in Medicine 2026

Abstract

Traditional radiology education is constrained by a restricted apprenticeship model and a scarcity of datasets structured for building artificial intelligence (AI)-based radiology education systems. To address this problem, we developed a novel end-to-end framework for transforming vast clinical archives into scalable radiology education resources. The proposed framework converts static radiographic data into an interactive learning system through three integrated components. First, a multi-stage curation pipeline establishes a foundation of trustworthy cases suitable for radiology education from noisy public archives. Second, a large language model pipeline automatically generates a rich library of questions engineered to build core radiology reasoning skills. Finally, this content is deployed on an interactive, gamified platform that uses an adaptive algorithm to deliver a personalized and engaging learning experience. The curation pipeline distilled an initial pool of 493,785 images into a final dataset of 881 high-fidelity chest radiographs, from which the automated content generation pipeline produced 2305 multiple-choice questions. The system was implemented as the League of Radiologists, a publicly accessible platform (https://radontology.org), demonstrating the feasibility of the proposed end-to-end architecture. A field demonstration resulted in 40 registered users and 68 unique examination sessions without technical failure, with 37.5% of active participants returning for multiple sessions. While currently focused on single finding chest radiographs, this study provides a practical and reproducible blueprint for implementing an AI-enabled adaptive radiology education platform using heterogeneous clinical imaging data. The described framework offers an extensible foundation for future development and evaluation of AI-driven educational systems in medical imaging.

Keywords Radiology education · Interactive learning · Artificial intelligence · End-to-End framework · Gamification

Introduction

The cornerstone of radiology education has traditionally been the apprenticeship model, in which trainees develop expertise through exposure to diverse cases under expert mentorship. However, this model is increasingly strained by rising imaging volumes, growing case complexity, and a shortage of expert educators, placing pressure on traditional training programs [1, 2]. As faculty balance escalating clinical demands with educational responsibilities, opportunities for structured case review, individualized feedback, and deliberate practice become limited [3, 4]. The result is a widening gap between traditional instruction and the needs of modern trainees, who require scalable and adaptive learning experiences to master the growing complexity of radiologic practice [5, 6].

Artificial intelligence (AI) presents a powerful solution to these challenges. Yet, despite promising applications — from generating exam-style questions to providing differential diagnoses [7–9] — its widespread integration into radiology curricula has stalled [10]. A central barrier is the scarcity of high-quality datasets curated specifically for pedagogy. Most large imaging datasets were created to benchmark AI models and therefore lack the instructional reliability needed for human training, as they often contain

Hyunji Kim, Young-Tak Kim and Saul Langarica contributed equally to this work.

Extended author information available on the last page of the article

Published online: 20 April 2026

label noise unsuitable for foundational learning [11, 12]. This data-centric challenge is compounded by a development focus on clinical tasks, like classification and report generation, rather than on integrating proven pedagogical strategies to sustain learner engagement [13]. This prioritization has yielded a collection of valuable but disconnected tools, not integrated educational ecosystems required for effective education.

To bridge these gaps, we introduce a novel end-to-end educational framework designed to systematically transform static clinical archives into trustworthy, curated datasets and dynamic, interactive learning resources. Our framework integrates three synergistic components: (1) A multi-stage curation process constructs a pedagogically reliable dataset through systematic label-noise reduction. (2) Automated content generation utilizes a large language model (LLM) pipeline to scalably create pedagogical questions aligned with core radiological reasoning categories. (3) Adaptive learner engagement delivers this content through an interactive, gamified platform that personalizes difficulty to foster sustained learning. This study details the pilot implementation of this framework specifically for chest radiography through the development of the 'League of Radiologists (LOR)' platform, providing a concrete example of how clinical imaging archives can be operationalized into an AI-enabled radiology education environment.

Methods

The proposed end-to-end framework was implemented as a pilot study through a systematic, three-stage process designed to transform heterogeneous public imaging archives into an operational AI-enabled radiology education system. First, a foundation of controlled input data was established through a multi-stage curation pipeline that integrates AI-based automated filtering with expert-defined inclusion criteria to identify single-finding chest radiograph cases for educational use. Second, the curated dataset was used to automatically generate a scalable library of multiple-choice questions aligned with core radiological reasoning categories. Finally, the generated content was deployed on an interactive, gamified platform engineered to provide a personalized and engaging learning experience.

Establishing a Foundation of Trustworthy Data

A pedagogically reliable dataset was constructed from an initial pool of 493,785 images across three public chest radiography archives through systematic label-noise reduction (Fig. 1). From these sources, we initially aggregated an expert-labeled cohort comprising 4596 cases with radiologist-provided annotations. Specifically, the PadChest-GR dataset provided 4555 studies with bounding box annotations for localization-based questions [14]. The MIMIC-CXR-JPG dataset (v2.0.0) offered 377,110 radiographs with free-text reports for report-driven question generation [15, 16]. Finally, the NIH ChestX-ray14 dataset provided 112,120 radiographs with structured disease labels [17].

The labels from the three sources were standardized into a unified set of seven thoracic pathologies: atelectasis, cardiomegaly, pleural effusion, infiltrate (opacity), mass, nodule, and pneumothorax. These categories were derived from the original NIH ChestX-ray8 label set [18], excluding pneumonia due to its high inter-observer variability and the necessity of clinical correlation for definitive diagnosis [19, 20].

To ensure the final dataset possessed both high pedagogical reliability and sufficient case volume, two complementary cohorts were assembled. The expert-labeled cohort served as a source of high-confidence cases with radiologist-provided annotations. It was composed of 3099 studies from PadChest-GR that were annotated through a multi-stage manual process [14], 687 test set studies from MIMIC-CXR-JPG annotated by a board-certified radiologist following the established CheXpert clinical labeling framework, ensuring reliable and standardized annotation [18], and the 810-image test set from NIH ChestX-ray14 labeled by a majority vote of five radiologists [21]. Only studies containing a single target finding among the seven categories were retained to reduce ambiguity in downstream automated processing.

Second, the AI-verified cohort was derived to expand case volume while preserving label precision. An automated filtering pipeline was applied to NIH ChestX-ray14, generating initial predictions using three publicly available pretrained classifiers based on the CheXNet architecture [22–24]. The resulting predictions were filtered by applying a confidence threshold specifically chosen to yield a 100% Positive Predictive Value (PPV) across the dataset, selecting only the cases where both the AI models and the original labels unanimously agreed on the presence of a finding. This quantitative step was followed by qualitative verification with saliency maps to ensure clinical plausibility. The strict PPV criterion was adopted to maximize label precision and minimize the risk of introducing false-positive cases, thereby ensuring that only the most reliable predictions were incorporated into the AI-verified cohort. Full technical details of this pipeline and diagnostic characteristics are provided in the Online Resource Methods, Online Resource Table 1, and Online Resource Fig. 1.

Automated Generation of Educational Content

An automated pipeline enabled by an LLM was developed to scalably generate educational questions from the curated dataset. This process was designed to produce a

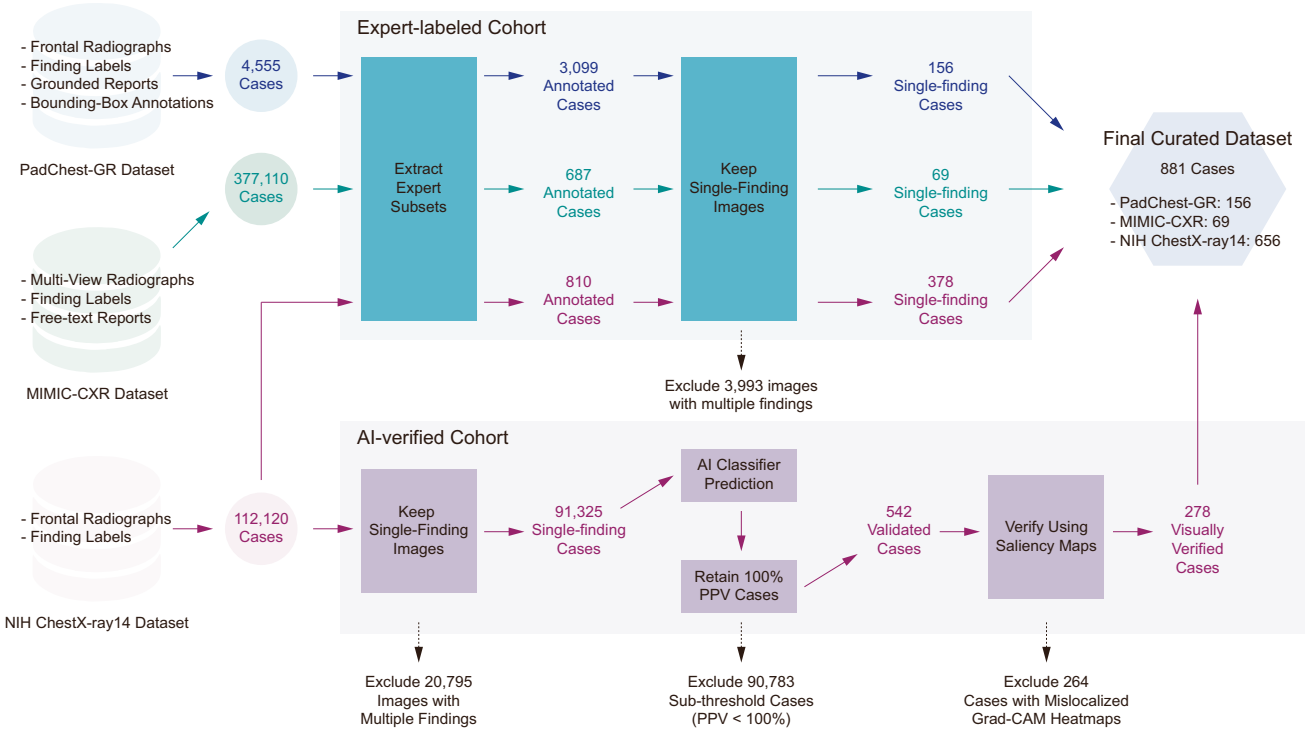

Fig. 1 Illustration of the Data-Curation Pipeline. Schematic representation of the multi-stage curation process for PadChest-GR, MIMIC-CXR, and NIH ChestX-ray14. The workflow illustrates the stratification into an Expert-labeled Cohort (top) and an AI-verified Cohort (bottom). Key components include the aggregation of expert-labeled subsets, the retention of single-finding studies (atelectasis, cardiomegaly, pleural effusion, infiltrate (opacity), mass, nodule, pneumothorax, and no finding), and the generation of an AI-verified cohort via ensemble classifiers (100% PPV threshold) and saliency map verification. The merger of these validated cohorts yields the final curated dataset ($n = 881$) used for question generation. *PPV* positive predictive value, *CXR* chest radiograph, *Grad-CAM* gradient-weighted class activation mapping

vast library of items targeting five core radiological reasoning categories which are finding detection, diagnosis, finding attributes, finding identification, and bounding box drawing (Table 1). The pipeline's architecture, illustrated in Fig. 2, consisted of three sequential modules: an Ontology Module for knowledge extraction, a Generation Module for question construction, and a Review Module for automated quality assurance.

Ontology Module

The Ontology Module processed heterogeneous input from the source datasets, including unstructured reports, categorical labels, and bounding box annotations, into a structured knowledge graph. For each report, the primary sentence describing the target finding was isolated, and attributes such as location and morphology were extracted. Each attribute

Table 1 Five question categories to support diversified diagnostic reasoning

Question Type	Description	Example Question
Finding Detection	Determine whether the radiograph shows no abnormality or contains a reportable finding	"Are there any reportable findings in this radiograph?"
Diagnosis	Identify the most clinically significant finding in the image	"What is the most significant finding visible in the image?"
Finding Attributes	Assess a specific attribute of the finding, such as its anatomical location or descriptive characteristics	"Where is the pleural effusion located?", "What shape best describes the finding atelectasis?"
Finding Identification	Select the correct finding shown within the highlighted bounding box	"Which finding is shown inside the bounding box?"
Bounding Box Drawing	Draw a bounding box around a specified finding in the image	"Please identify the location of the atelectasis on the image."

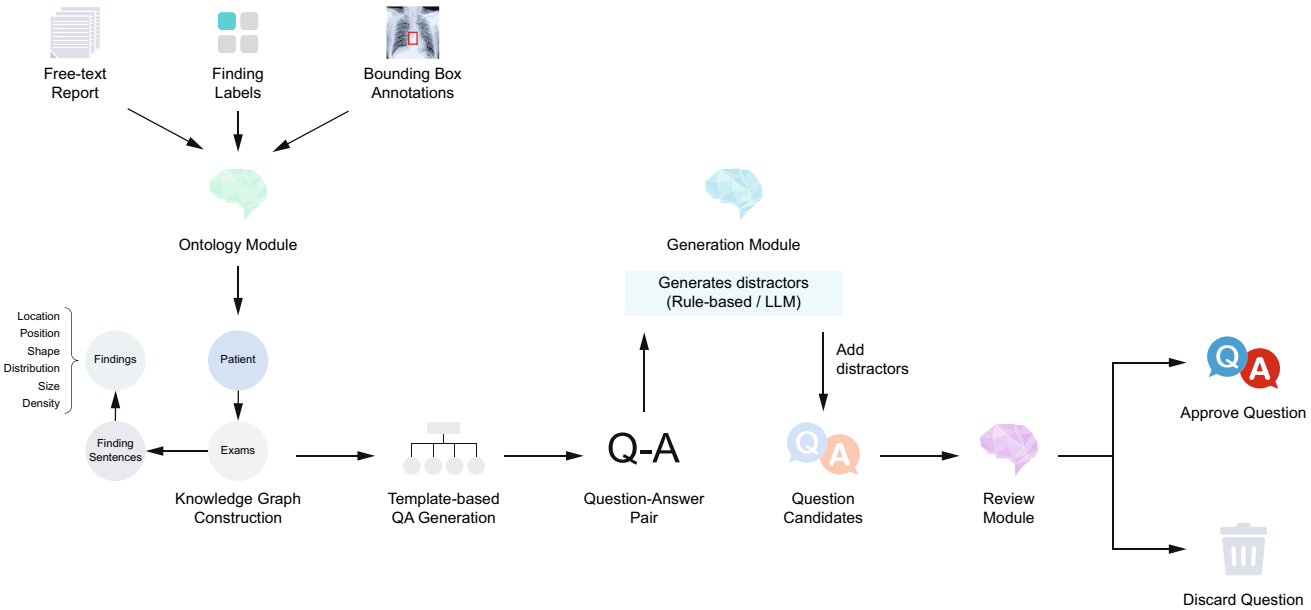

Fig. 2 Educational Question Generation Pipeline. The Ontology Module extracts structured attributes from free-text reports, finding labels, and bounding box annotations to construct a clinical knowledge graph. The Generation Module then produces question–answer pairs via template-based mapping and distractor generation. The Review Module provides final automated quality assurance for clinical relevance and clarity to determine item approval. *QA* question and answer, *LLM* large language model

underwent verification to confirm grounding in the source text. Validated attributes, together with structured labels and bounding boxes, were encoded into the knowledge graph for downstream question generation.

Generation Module

The Generation Module constructed multiple-choice questions from the structured output of the Ontology Module. Question stems and correct answers were generated by mapping attributes to predefined templates. Distractors were then produced by either sampling from predefined pathology lists for discrete-answer categories or generating alternatives to ensure plausibility and clinical coherence for descriptive categories.

Review Module

The Review Module automated LLM driven quality assurance. Guided by a prompt establishing the persona of an academic thoracic radiologist, the model was instructed to follow a structured JSON schema to assign style scores, categorize clinical training levels, and verify diagnostic fidelity to the source text. Each question and answer set was evaluated against predefined criteria including clinical relevance, linguistic clarity, and estimated difficulty. Based on this comprehensive assessment, items were either accepted, automatically revised if salvageable, or rejected if they failed to meet the required standards.

All generative steps used Med-Gemma-4B [25], a model adapted for the medical domain, executed locally on an NVIDIA DGX system with four Tesla V100-SXM2 GPUs (32 GB each, Python 3.12). This model was selected to balance domain adaptation and computational efficiency, enabling reproducible generation under commonly available hardware constraints. Detailed prompt specifications are provided in the Online Resource Method 2.

Deployment within an Interactive Learning Environment

The curated educational content was implemented within a web-based learning platform designed to provide an interactive and personalized learning experience (Fig. 3). The platform was built around three core components: an adaptive learning algorithm to personalize difficulty, gamification features to enhance learner engagement, and a continuous quality assurance system to maintain content integrity over time.

Gamification and Engagement Features

To enhance learner engagement, the platform incorporates gamification features designed to stimulate motivation and consistent practice. A typical session follows a high-tension survival format consisting of ten dynamic questions. Users begin with a strict 120-s global countdown timer. To reward diagnostic speed, the platform provides a 10-s time extension and an instantaneous success sound effect for every

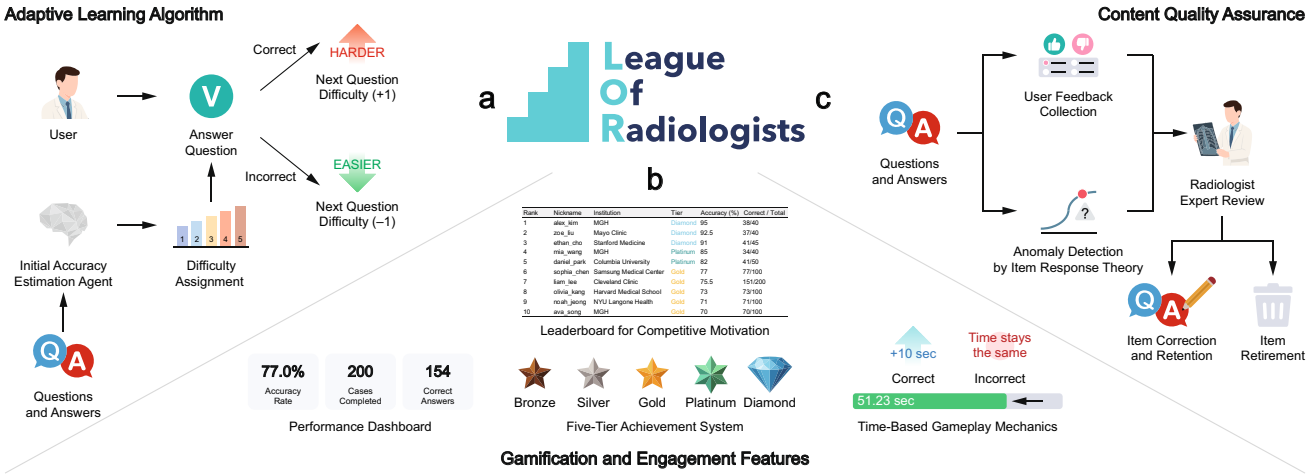

Fig. 3 Workflow of the League of Radiologists (LOR) platform. **a** Adaptive learning algorithm calibrates question difficulty dynamically based on individual learner performance. **b** Gamification and engagement features incorporate a performance dashboard, a five-tier achievement system, and time-based gameplay mechanics. **c** Content quality assurance combines user feedback, statistical anomaly detection, and radiologist expert review to refine or retire question items

correct answer. The current level of a user is calculated using the following equation

$$Level\ of\ a\ user = \lfloor 1 + \log_2(Cumulative\ Correct\ Answers + 1) \rfloor$$

Achievement badges and tiers are awarded only after a user completes a minimum of ten questions to maintain statistical reliability. Learners are assigned to one of five tiers including Bronze, Silver, Gold, Platinum, or Diamond based on their cumulative accuracy. Rankings are updated in real-time via a global leaderboard that prioritizes the achieved tier first and cumulative accuracy as a secondary sorting factor.

Adaptive Learning Algorithm

The platform utilizes a dual-layered adaptive difficulty algorithm to maintain learners within their optimal zone of proximal development. At the longitudinal level, the system calibrates the initial difficulty of each session based on the cumulative tier of the learner. For example, a Diamond tier user initiates a session with questions explicitly drawn from the Level 5 difficulty bin which contains items with the lowest accuracy percentiles. Conversely, a beginner in the Bronze tier begins with foundational Level 1 questions. At the session level, the inherent difficulty of subsequent items is dynamically adjusted in real time based on immediate diagnostic performance. To establish an objective difficulty mapping, each question is assigned a baseline accuracy score of 0.5 upon generation. This parameter is iteratively updated utilizing a Bayesian smoothing formula as historical user data accumulate. The calculation incorporates the aggregate of correct responses and total attempts to dynamically reclassify items into five distinct difficulty tiers. During active gameplay, a correct diagnosis increments the difficulty tier of the succeeding question by one level, whereas an incorrect response decrements the tier by one level to ensure a continuous optimal flow state (Fig. 4).

Content Quality Assurance

A content quality assurance system is integrated to enable post-deployment monitoring of item-level performance (Fig. 5). To maintain uninterrupted system performance during real-time user interaction, this quality control mechanism utilizes an asynchronous processing pipeline. Direct subjective feedback measured by a 5-point Likert scale is logged instantly during active gameplay. Concurrently, intensive statistical evaluations are executed in the background. For initial screening of items with fewer than 200 responses, the system uses classical test theory (CTT) to detect issues like negative discrimination [26]. After sufficient data is collected, item response theory (IRT) is applied to identify more nuanced item failures using misfit statistics. Once an item is flagged by either user feedback or these primary statistical metrics, the system calculates the impact of its permanent deletion on the overall internal consistency (Cronbach's alpha). Based on this dual-validation fail-safe, flagged items are then either automatically removed from rotation or retained for subsequent manual expert review, revision, or retirement.

The system utilizes Python 3.9 with the Streamlit 1.40 framework to serve as both the unified frontend interface and backend router. Data handling is structurally partitioned where a Neo4j version 5 graph database manages the storage and retrieval of educational content. Furthermore, user authentication and basic administrative data are securely handled by a SQLite engine. The entire application is deployed on an Amazon Web Services EC2 instance to provide stable access for global users.

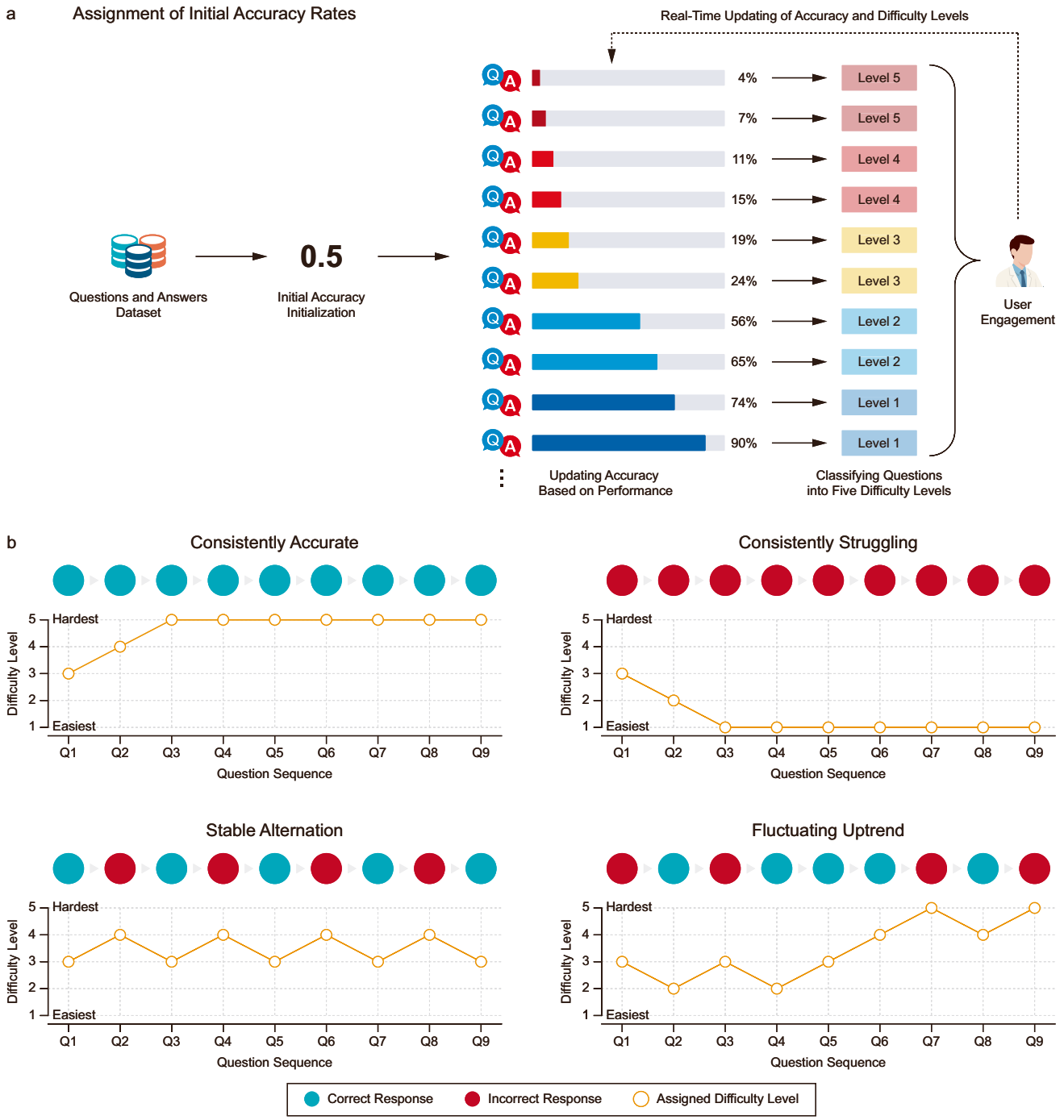

Fig. 4 Adaptive learning algorithm in the League of Radiologists (LOR) platform. **a** Initialization phase assigns baseline accuracy rates to question items and categorizes them into five difficulty tiers (Levels 1–5) based on accumulated performance data. **b** Real-time adaptive adjustment dynamically modifies question difficulty during quiz sessions, with the level advancing or regressing based on individual response accuracy. Representative trajectories across a nine-question session are shown

Results

Pedagogically Curated Dataset

The curation pipeline distilled an initial pool of 493,785 images into a final dataset of 881 single-finding chest radiographs (Fig. 1). The expert-labeled cohort ($n=603$) was derived from 4596 radiologist-annotated studies across three datasets, with a 13.1% acceptance rate representing the proportion of initial studies that satisfy all inclusion, filtering, and quality control

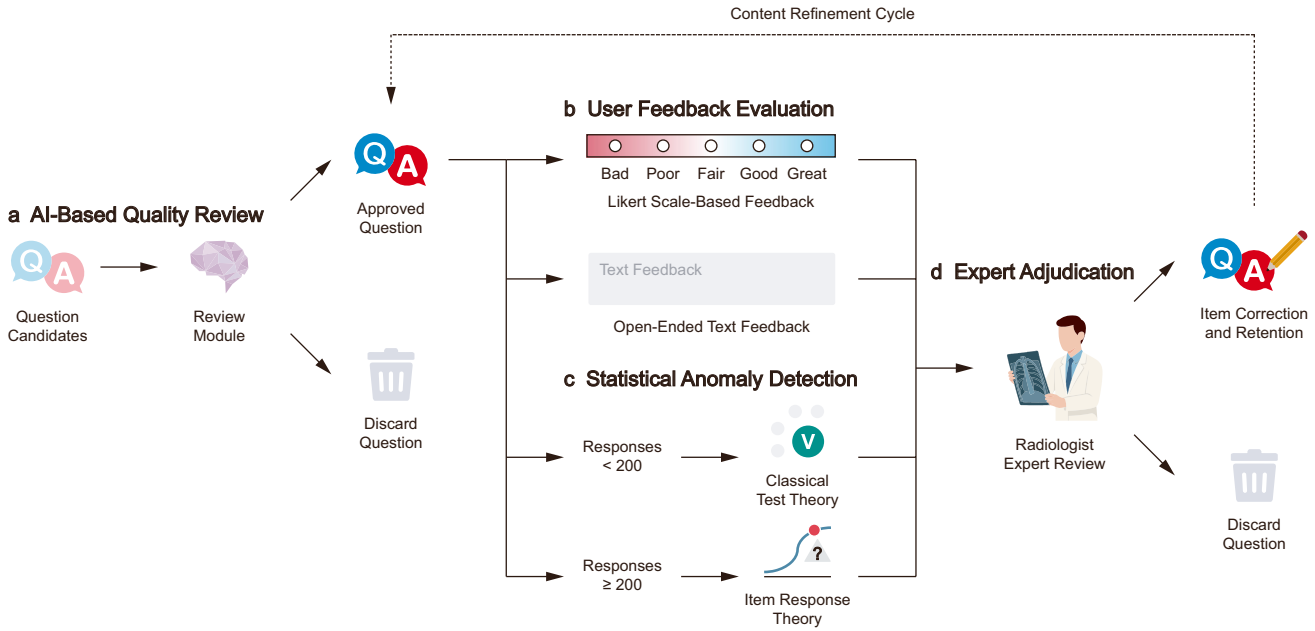

Fig. 5 Multi-step content quality management and validation workflow in the League of Radiologists (LOR) platform. **a** An AI-based Review module performs an initial quality screening of all question candidates. **b** Approved items undergo user feedback evaluation via Likert scale ratings and open-ended comments. **c** Statistical anomaly detection utilizing classical test theory for items with fewer than 200 responses and item response theory for those with 200 or more, targeting issues such as negative discrimination or misfit statistics. **d** Radiologist expert adjudication of flagged items to determine revision and retention through the content refinement cycle or permanent discard

criteria. The AI-verified cohort ($n=278$) was obtained by applying a filtering pipeline to 91,325 single-finding images from the NIH ChestX-ray14 dataset. This process included quantitative filtering using a 100% positive predictive value threshold ($n=542$), followed by saliency map–based screening, yielding a final acceptance rate of 0.3%. From the curated dataset, the automated content generation pipeline produced 2305 multiple-choice questions across five core radiological reasoning categories and seven common thoracic pathologies (Table 2).

Implementation of an Interactive Learning Platform and User Journey

The curated dataset and generated educational content were deployed within a publicly accessible, web-based platform, available at https://radontology.org. The platform was implemented as an interactive learning system that integrates structured learning modes, adaptive learning design, and gamification mechanisms to operationalize the proposed end-to-end framework. Upon authentication, users access a centralized Dashboard that serves as the primary navigation hub. From this interface, participants navigate between distinct educational pathways which include the Tutorial Mode, the Study Mode, and the Main Game Mode. The Tutorial Mode functions as an introductory environment. The Study Mode offers a self-paced educational experience where learners select specific radiological pathologies, review foundational Radiology Insights, and complete a tailored quiz without time constraints. The Main Game Mode constitutes the core competitive environment. To preserve

Table 2 Distribution of generated questions by reasoning category and pathology

Reasoning Category	Atelectasis	Cardiomegaly	Mass	Nodule	Pleural Effusion	Pneumothorax	Infiltrate (Opacity)	No Finding
Finding Detection	103	242	25	61	101	25	64	260
Diagnosis	103	242	25	61	101	25	64	260
Finding Attributes	25	6	3	49	60	12	95	0
Finding Identification	5	62	2	24	0	2	42	0
Bounding Box Drawing	5	62	2	24	19	2	42	0

immersive gameplay, pedagogical feedback and detailed clinical justifications are deferred until the completion of a full diagnostic session. Representative user interface views from the deployed platform are shown in Fig. 6, demonstrating the practical implementation of the proposed framework in a live system. Key features of the user interface, including learning modes, adaptive tier progression, gamification elements, and the quality assurance workflow, are shown in Online Resource Figs. 2–5. A video overview of the LOR platform, illustrating its core features and user interface, is available at Online Resource Video.

Field Demonstration and Pilot Deployment

The LOR platform was deployed as a free educational platform at the Society for Imaging Informatics in Medicine (SIIM) 2025 Annual Meeting AI Playground. An early version of the curated dataset was used to power the platform during the three-day event from May 21 to May 23 in 2025. The promotion of the platform resulted in 40 registered users, of whom 24 actively engaged with the system across multiple sessions. The framework demonstrated its operational scalability by handling a total of 68 unique examination sessions without technical failure. Among the users who reported their professional experience ($n = 19$), 11 individuals had less than one year of radiology experience and 6 had over four years, demonstrating the ability of the system to engage a diverse audience. The platform further demonstrated its usability as 9 users representing 37.5% of the active base participated in two or more sessions with one individual completing 9 distinct attempts.

The deployed version incorporates expert feedback collected during pilot evaluation, with representative before–after refinements summarized in Online Resource Table 2.

Discussion

This study introduces an end-to-end system for transforming large public clinical imaging archives into an operational radiology education platform. To validate this architecture, we implemented a pilot system focused on chest radiography. The proposed framework integrates three core components: (i) curation of trustworthy cases through a multi-stage, AI-assisted curation process, (ii) automated generation of educational content using an LLM-based pipeline, and (iii) deployment of this content within an interactive, gamified platform. This work focuses on describing a reproducible technical pathway for converting heterogeneous clinical data into structured educational artifacts. The framework provides a foundation for future

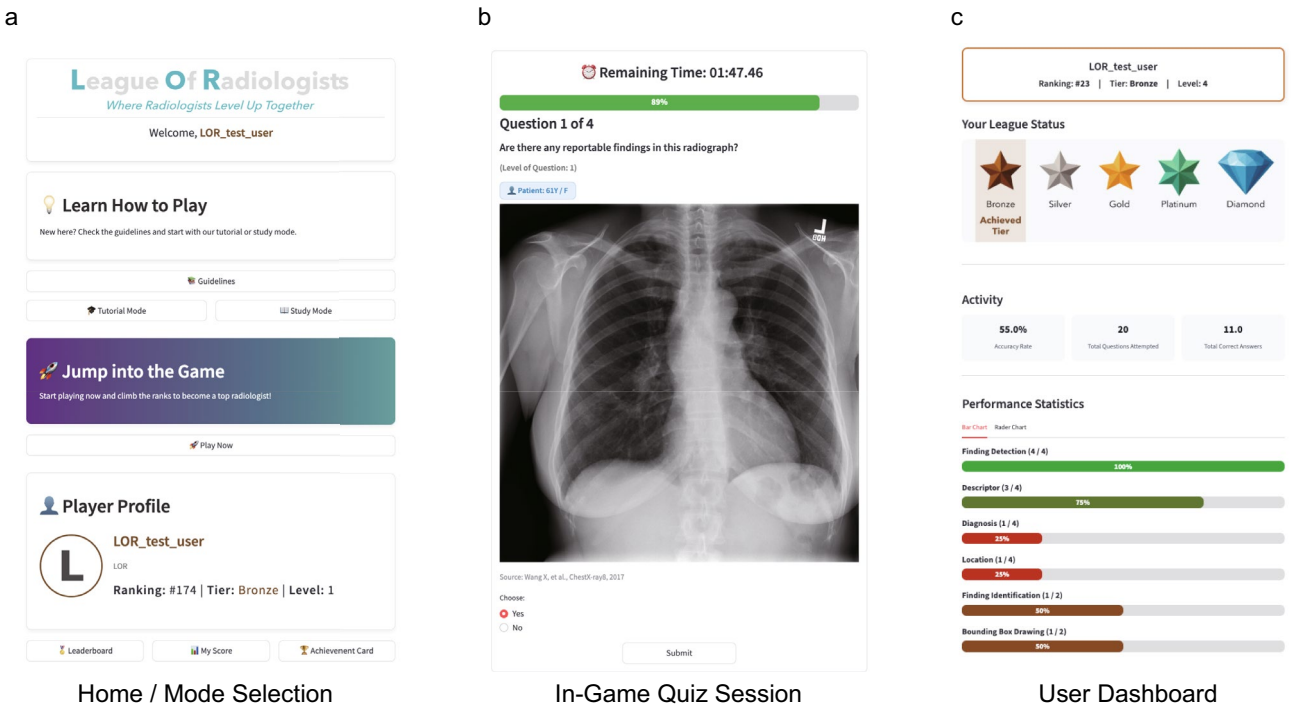

Fig. 6 Representative user interface panels from the League of Radiologists (LOR) platform. **a** Home interface displaying access to core platform functions, including Tutorial mode, Study mode, Main Game mode, and user guidelines. **b** Time-limited quiz sessions incorporating anonymized patient demographics, such as age and gender, to facilitate clinical contextualization within a simulated test environment. **c** User Dashboard summarizing overall accuracy, attempted questions, and category-specific performance

AI-enabled radiology education systems and downstream educational evaluation studies across broader medical imaging domains.

A central contribution of this work is the explicit focus on data preparation and content structuring for educational use. Previous radiology-oriented datasets were primarily designed for algorithm benchmarking rather than for instructing trainees [27]. Their pedagogical value is often compromised by label noise and the inclusion of diagnostically complex, multi-finding cases unsuitable for foundational learning. Our curation pipeline was designed to support controlled system inputs by restricting inclusion to single-finding radiographs and by combining expert-labeled cases with an AI-verified cohort generated through predefined filtering criteria. Furthermore, the generated questions were mapped to five core reasoning categories, ensuring the content aligned with established radiology training curricula [28–30]. Together, these design choices limited downstream automated content generation to datasets curated to reduce diagnostic ambiguity, a key consideration for educational applications.

While high-quality curated data forms the foundation of effective medical training, pedagogical delivery mechanisms in radiology education face distinct systemic challenges. Traditional learning modalities such as static online modules, flipped classrooms, and case-based platforms provide excellent foundational knowledge but frequently struggle to offer scalable and real-time individualized feedback [3, 4, 13]. Recent literature emphasizes the transition toward precision education, highlighting the critical need to transform medical training by dynamically tailoring instruction to individual learner needs [31]. The proposed AI-based framework directly reflects these contemporary requirements by continuously adapting task difficulty and providing instantaneous clinical justifications. Unlike flipped classrooms that demand intensive faculty resources or static modules that rely on standardized progression, the developed platform automates individualized guidance. This architectural shift effectively overcomes the scalability and personalization bottlenecks inherent to conventional educational modalities (Table 3).

Label noise in large public datasets remains a major barrier to the use of AI in medical education, as it undermines the clinical validity of derived content [32]. We addressed this challenge with a multi-layered validation strategy. The dataset construction combined expert-labeled cases from publicly available radiologist-annotated subsets with an AI-verified cohort selected using a strict positive predictive value threshold and saliency map–based screening. These steps were designed to constrain downstream content generation to image–label pairs with reduced ambiguity. This process was strengthened by expert feedback, in which board-certified radiologists reviewed an early version of the dataset and provided feedback that refined the generation pipeline, thereby enhancing both clinical and educational integrity (Online Resource Result 1 and Online Resource Table 2). In addition, post-deployment monitoring mechanisms were implemented within the platform to flag items for further review based on user feedback signals and statistical indicators (Online Resource Result 2 and Online Resource Fig. 5). Together, these mechanisms illustrate how label noise can be systematically managed across data preparation, content generation, and deployment stages in an AI-enabled educational system.

The pedagogical design of the LOR system was intentionally developed to overcome key limitations of traditional radiology training, particularly the lack of personalized feedback and difficulty sustaining learner engagement at scale [5, 6]. The adaptive learning algorithm directly addresses the need for individualized learning paths by adjusting question difficulty in real time. In parallel, the integration of gamification mechanics enhances motivation and knowledge retention, strategies that have demonstrated effectiveness in medical education [3, 4]. This pedagogical framework is actively reinforced by its content design. A focus on single-finding cases builds a foundation for mastering core concepts, while five structured question categories guide the learner through the entire radiological reasoning process, from perception to interpretation. This design augments traditional training by providing a scalable resource for engaging deliberate practice.

This study has several limitations, and these inform directions for future work. First, the proposed framework relies on preexisting textual data, such as structured labels and free-text reports, restricting its direct applicability to completely raw or unlabeled clinical archives. Additionally, the cases curated for the educational resource were intentionally restricted to single, unambiguous findings to create a controlled learning environment and therefore do not encompass the multi-finding scenarios common in clinical practice. Addressing these limitations will require integrating automated image-to-text or computer vision modules to process raw archives, alongside expanding the

Table 3 Comparison of existing radiology education modalities and the proposed AI-based platform

Educational Modality	Content Generation	Difficulty Calibration	Feedback Mechanism
Static Online Modules	Manual and static content	Static	Delayed or generic
Flipped Classrooms	Faculty-prepared materials	Session-level	Real time but faculty dependent
Case Based Platforms	Expert-curated case libraries	User paced	Limited or generic explanations
Proposed Framework	Scalable (LLM-assisted generation)	Dynamically adaptive	Real time and individualized

framework to include multi-finding cases and developing question designs that capture diagnostic reasoning in more complex clinical contexts. Despite the use of a multi-stage quality assurance pipeline, LLM-based generation still carries the risk of subtle inaccuracies such as temporal context confusion or inconsistent difficulty calibration that may not be fully identified through automated screening. Future efforts should incorporate the integration of chain of thought reasoning to LLMs alongside broader expert review cycles and cross-institutional validation to further safeguard content reliability. In addition, the current work is a pilot limited to chest radiography, and further research is needed to extend the framework to other two-dimensional modalities such as musculoskeletal radiographs or to redesign modules for volumetric modalities such as CT to handle three-dimensional data complexity. Finally, this work focused on system design and implementation rather than prospective educational evaluation. Formal studies utilizing pre-test and post-test methodologies are required to quantitatively assess key learner outcomes including diagnostic accuracy, long-term knowledge retention, and learner perceptions assessed through structured surveys. These future evaluations will target diverse trainee populations including medical students, radiology residents, and clinical fellows.

Conclusion

This study proposes an end-to-end framework that systematically transforms public clinical archives into reproducible, pedagogically reliable educational resources. By combining expert-labeled and AI-verified cases, our curation pipeline addresses the persistent challenge of label noise, while an LLM enables scalable generation of curriculum-aligned questions. These resources are delivered through an interactive, gamified learning platform that personalizes difficulty and sustains engagement. Together, this framework provides a reproducible pathway for converting static imaging archives into an operational radiology education tool, offering a scalable solution to the shortage of trustworthy training resources in radiology education. Its proof-of-concept implementation for chest radiography demonstrates the technical feasibility of the proposed framework and its potential for broader application in AI-enabled medical education across various radiological domains. Ultimately, this scalable architecture establishes a reproducible pathway for AI-enabled medical education and the digital transformation of clinical archives.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10278-026-01960-w.

Acknowledgements This study was partially supported by Amazon Web Services (AWS). The LOR platform was demonstrated at the Society for Imaging Informatics in Medicine (SIIM) 2025 Annual Meeting AI Playground. The authors thank the National Institutes of Health Clinical Center for providing the ChestX-ray14 dataset, and the creators of the PadChest-GR dataset for their meticulously annotated data. We also acknowledge the National Institutes of Health for making the MIMIC-CXR-JPG (v2.0.0) dataset available on PhysioNet. We are deeply grateful to the original authors of all datasets for their significant contributions to the research community. Finally, we extend our sincere appreciation to the board-certified radiologists who provided critical feedback during the development process and for their role in the continuous quality assurance of the LOR platform.

Author Contribution All authors contributed to the study conception and design. Study conception and design were performed by H. Kim, Y–T. Kim, S. Langarica, and S. Do. Material preparation, data acquisition, platform development, and analysis were performed by H. Kim, Y–T. Kim, and S. Langarica. The first draft of the manuscript was written by H. Kim and Y–T. Kim. Clinical validation and manuscript revision were performed by S. Langarica, K.P. Fialkowski, J.C.Y. Seah, J.S.N. Tang, K.D. Song, D.C. Jung, K.T. Bae, R.L. Cochran, M.D. Succi, S. McDermott, M. Bahl, J.B. Ackman, M.H. Lev, M.S. Gee, and S. Do. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This study was partially supported by Amazon Web Services (AWS).

Data Availability The imaging datasets used in this study are publicly available open datasets. The data generated during this study are available from the corresponding author upon request.

Declarations

Ethics Approval This study was reviewed by the MGB Institutional Review Board (Protocol #2025P002869) and determined to be exempt under 45 CFR 46.104(d)(1)(2). The study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments.

Consent to Participate The requirement for informed consent was waived by the MGB Institutional Review Board due to the retrospective nature of the study.

Consent for Publication Not applicable as the study involves the use of anonymized data and does not contain identifiable details of any individual.

Conflict of interest The authors declare no competing interests.

References

1. Oeppen RS, Rafiee H, Suresh P, Rajesh A, Parekh A: Navigating the challenges in radiology training expansion: costs and benefits. Clin Radiol 81:106767, 2025
2. Aggarwal A, Lazarow F, Anzai Y, Elsayed M, Ghobadi C, Dandan OA, Griffith B, Straus CM, Kadom N: Maximizing Value While Volumes are Increasing. Curr Probl Diagn Radiol 50:451–453, 2021
3. Ge L, Chen Y, Yan C, Chen Z, Liu J: Effectiveness of flipped classroom vs traditional lectures in radiology education: A meta-analysis. Medicine (Baltimore) 99:e22430, 2020
4. Stirrat T, Martin R, Umair M, Waller J: Advancing radiology education for medical students: leveraging digital tools and resources. Pol J Radiol 89:e508–e516, 2024
5. Afshari Mirak S, Tirumani SH, Ramaiya N, Mohamed I: The Growing Nationwide Radiologist Shortage: Current

Opportunities and Ongoing Challenges for International Medical Graduate Radiologists. Radiology 314:e232625, 2025
6. Jing AB, Garg N, Zhang J, Brown JJ: AI solutions to the radiology workforce shortage. Npj Health Syst 2:20, 2025
7. Shah C, Davtyan K, Nasrallah I, Bryan RN, Mohan S: Artificial Intelligence-Powered Clinical Decision Support and Simulation Platform for Radiology Trainee Education. J Digit Imaging 36:11–16, 2023
8. Lyo S, Mohan S, Hassankhani A, Noor A, Dako F, Cook T: From Revisions to Insights: Converting Radiology Report Revisions into Actionable Educational Feedback Using Generative AI Models. J Imaging Inform Med 38:1265–1279, 2025
9. Borgbjerg J, Thompson JD, Salte IM, Frøkjær JB: Towards AI-augmented radiology education: a web-based application for perception training in chest X-ray nodule detection. Br J Radiol 96:20230299, 2023
10. Tejani AS, Elhalawani H, Moy L, Kohli M, Kahn CE: Artificial Intelligence and Radiology Education. Radiol Artif Intell 5:e220084, 2023
11. Valikodath NG, Cole E, Ting DSW, Campbell JP, Pasquale LR, Chiang MF, Chan RVP, American Academy of Ophthalmology Task Force on Artificial Intelligence: Impact of Artificial Intelligence on Medical Education in Ophthalmology. Transl Vis Sci Technol 10:14, 2021
12. Lee J, Wu AS, Li D, Kulasegaram K (Mahan): Artificial Intelligence in Undergraduate Medical Education: A Scoping Review. Acad Med 96:S62, 2021
13. Awan O, Dey C, Salts H, Brian J, Fotos J, Royston E, Braileanu M, Ghobadi E, Powell J, Chung C, Auffermann W: Making Learning Fun: Gaming in Radiology Education. Acad Radiol 26:1127–1136, 2019
14. de Castro DC, Bustos A, Bannur S, Hyland SL, Bouzid K, Wetscherek MT, Sánchez-Valverde MD, Jaques-Pérez L, Pérez-Rodríguez L, Takeda K, Salinas-Serrano JM, Alvarez-Valle J, Galant-Herrero J, Pertusa A: PadChest-GR: A Bilingual Chest X-Ray Dataset for Grounded Radiology Report Generation. NEJM AI 2:AIdbp2401120, 2025
15. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C, Mark RG, Horng S: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6:317, 2019
16. Johnson A, Pollard T, Mark R, Berkowitz S, Horng S: MIMIC-CXR Database. PhysioNet, https://doi.org/10.13026/4jqj-jw95, July 23, 2024
17. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2097–2106, 2017
18. Johnson A, Lungren M, Peng Y, Lu Z, Mark R, Berkowitz S, Horng S: MIMIC-CXR-JPG: chest radiographs with structured labels. PhysioNet, https://doi.org/10.13026/jsn5-t979, March 12, 2024
19. Franquet T: Imaging of Community-acquired Pneumonia. J Thorac Imaging 33:282–294, 2018
20. Self WH, Courtney DM, McNaughton CD, Wunderink RG, Kline JA: High discordance of chest x-ray and computed tomography for detection of pulmonary opacities in ED patients: implications for diagnosing pneumonia. Am J Emerg Med 31:401–405, 2013
21. Nabulsi Z, Sellergren A, Jamshy S, Lau C, Santos E, Kiraly AP, Ye W, Yang J, Pilgrim R, Kazemzadeh S, Yu J, Kalidindi SR, Etemadi M, Garcia-Vicente F, Melnick D, Corrado GS, Peng L, Eswaran K, Tse D, Beladia N, Liu Y, Chen P-HC, Shetty S: Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and COVID-19. Sci Rep 11:15523, 2021
22. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP, Ng AY: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:1711.05225, 2017
23. Zech J: reproduce-chexnet. Available at https://github.com/jrzech/reproduce-chexnet. Accessed 5 February 2026.
24. Strick D, Garcia C, Huang A: Reproducing and Improving CheXNet: Deep Learning for Chest X-ray Disease Classification. arXiv:2505.06646, 2025
25. Sellergren A, Kazemzadeh S, Jaroensri T, Kiraly A, Traverse M, Kohlberger T, Xu S, Jamil F, Hughes C, Lau C, Chen J, Mahvar F, Yatziv L, Chen T, Sterling B, Baby SA, Baby SM, Lai J, Schmidgall S, Yang L, Chen K, Bjornsson P, Reddy S, Brush R, Philbrick K, Asiedu M, Mezerreg I, Hu H, Yang H, Tiwari R, Jansen S, Singh P, Liu Y, Azizi S, Kamath A, Ferret J, Pathak S, Vieillard N, Merhej R, Perrin S, Matejovicova T, Ramé A, Riviere M, Rouillard L, Mesnard T, Cideron G, Grill J, Ramos S, Yvinec E, Casbon M, Buchatskaya E, Alayrac J-B, Lepikhin D, Feinberg V, Borgeaud S, Andreev A, Hardin C, Dadashi R, Hussenot L, Joulin A, Bachem O, Matias Y, Chou K, Hassidim A, Goel K, Farabet C, Barral J, Warkentin T, Shlens J, Fleet D, Cotruta V, Sanseviero O, Martins G, Kirk P, Rao A, Shetty S, Steiner DF, Kirmizibayrak C, Pilgrim R, Golden D, Yang L: MedGemma Technical Report. arXiv:2507.05201, 2025
26. De Champlain AF: A primer on classical test theory and item response theory for assessments in medical education. Med Educ 44:109–117, 2010
27. Lin Z, Zhang D, Tao Q, Shi D, Haffari G, Wu Q, He M, Ge Z: Medical visual question answering: A survey. Artif Intell Med 143:102611, 2023
28. Cook TS, Samples M, Krishnaraj A: Patient- and Family-Centered Care in Radiology: Lessons Learned and Next Steps. J Am Coll Radiol 21:5–6, 2024
29. Waite S, Farooq Z, Grigorian A, Sistrom C, Kolla S, Mancuso A, Martinez-Conde S, Alexander RG, Kantor A, Macknik SL: A Review of Perceptual Expertise in Radiology-How it develops, How we can test it, and Why humans still matter in the era of Artificial Intelligence. Acad Radiol 27:26–38, 2020
30. Branstetter BF, Faix LE, Humphrey AL, Schumann JB: Preclinical Medical Student Training in Radiology: The Effect of Early Exposure. AJR Am J Roentgenol 188:W9–W14, 2007
31. Desai SV, Khan S, Lomis K: AI-Enabled Precision-Education Systems — Transforming Lifelong Learning in Medicine. N Engl J Med 394:838–841, 2026
32. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY: CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proceedings of the AAAI conference on artificial intelligence 33:590–597, 2019

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Hyunji Kim[1] · Young-Tak Kim[1] · Saul Langarica[2] · Kevin P. Fialkowski[1] · Jarrel C. Y. Seah[1] · Jennifer S. N. Tang[3] · Kyoung Doo Song[4] · Dae Chul Jung[5] · Kyongtae Tyler Bae[6] · Rory L. Cochran[1] · Marc D. Succi[1,7,8] · Shaunagh McDermott[1] · Manisha Bahl[1] · Jeanne B. Ackman[1] · Michael H. Lev[1] · Michael S. Gee[1] · Synho Do[1,9]

✉ Synho Do
sdo@mgh.harvard.edu

1 Department of Radiology, Massachusetts General Hospital, Harvard Medical School, 55 Fruit Street and 125 Nashua Street, Boston, MA 02114, USA

2 Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Marcoleta 367, Santiago, Región Metropolitana 8320165, Chile

3 Department of Radiology, St. Vincent's Hospital Melbourne, 41 Victoria Parade, Fitzroy, VIC, Victoria 3065, Australia

4 Department of Radiology, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-Ro, Gangnam-Gu, Seoul 06351, Republic of Korea

5 Department of Radiology, Severance Hospital, Yonsei University College of Medicine, 50-1 Yonsei-Ro, Seodaemun-Gu, Seoul 03722, Republic of Korea

6 Department of Diagnostic Radiology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Pok Fu Lam, Hong Kong SAR, China

7 Medically Engineered Solutions in Healthcare Incubator (MESH IO), Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA

8 Mass General Brigham Innovation, Mass General Brigham, 399 Revolution Drive, Somerville, MA 02145, USA

9 Kempner Institute, Harvard University, 150 Western Avenue, Boston, MA 02134, USA