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

CXR Insight Lab: An Auditable Multimodal AI Workbench for Transparent Chest X-Ray Interpretation

Seema Singh Saharan, PhD, UCSF; Rachel Johansson, MD; Iris Liu, MD; Beck Olson; Rajesh Shah, MD

Briefly describe what attendees will see during your demo.

During the live demonstration, attendees will experience the complete end-to-end workflow of CXR Insight Lab, from image ingestion to transparent report generation and audit validation. Beginning with a publicly available chest radiograph from the VinBigData dataset, the platform standardizes the image, extracts multimodal visual evidence using BiomedCLIP, generates a structured radiology report using Qwen3-VL, and presents every intermediate processing stage through an interactive browser interface. Rather than displaying only the final AI-generated report, attendees will see how evidence is linked to reported findings, how structured outputs are produced, and how deterministic validation evaluates the workflow before results are exported. The demonstration will showcase schema validation, evidence consistency checks, contradiction detection, unsupported finding identification, uncertainty assessment, and workflow disposition assignment. Visual overlays, structured JSON reports, audit logs, and machine-readable summaries will illustrate how each component contributes to transparent AI quality assurance. The demo will also include configurable failure scenarios that intentionally introduce unsupported findings, contradictory statements, and incomplete outputs to demonstrate how the validation engine identifies potential issues before report export. Finally, attendees will see how the platform's modular architecture enables rapid substitution and comparison of alternative vision-language models, detection algorithms, segmentation methods, and safety policies without changing the overall workflow. The demonstration emphasizes transparency, reproducibility, and responsible evaluation of multimodal AI systems, illustrating how auditable workflows can support research, education, benchmarking, and future clinical translation rather than functioning as a standalone diagnostic application.

During the live demonstration, attendees will experience the complete end-to-end workflow of CXR Insight Lab, from image ingestion to transparent report generation and audit validation. Beginning with a publicly available chest radiograph from the VinBigData dataset, the platform standardizes the image, extracts multimodal visual evidence using BiomedCLIP, generates a structured radiology report using Qwen3-VL, and presents every intermediate processing stage through an interactive browser interface. Rather than displaying only the final AI-generated report, attendees will see how evidence is linked to reported findings, how structured outputs are produced, and how deterministic validation evaluates the workflow before results are exported. The demonstration will showcase schema validation, evidence consistency checks, contradiction detection, unsupported finding identification, uncertainty assessment, and workflow disposition assignment. Visual overlays, structured JSON reports, audit logs, and machine-readable summaries will illustrate how each component contributes to transparent AI quality assurance. The demo will also include configurable failure scenarios that intentionally introduce unsupported findings, contradictory statements, and incomplete outputs to demonstrate how the validation engine identifies potential issues before report export. Finally, attendees will see how the platform's modular architecture enables rapid substitution and comparison of alternative vision-language models, detection algorithms, segmentation methods, and safety policies without changing the overall workflow. The demonstration emphasizes transparency, reproducibility, and responsible

evaluation of multimodal AI systems, illustrating how auditable workflows can support research, education, benchmarking, and future clinical translation rather than functioning as a standalone diagnostic application.

Problem & Motivation: What clinical or operational problem does your MVP address?

Chest radiography is the most frequently performed medical imaging examination worldwide, yet most artificial intelligence systems developed for chest X-rays remain narrowly focused on individual tasks such as disease classification, object detection, segmentation, or automated report generation. While these models often demonstrate impressive performance on benchmark datasets, they typically expose only the final prediction, providing little visibility into the intermediate reasoning, supporting image evidence, or potential failure points. As multimodal foundation models become increasingly integrated into radiology research, the lack of transparency, auditability, and systematic validation presents a significant barrier to trustworthy adoption, quality assurance, and clinical translation. Radiologists, AI developers, educators, and healthcare organizations increasingly need tools to inspect how AI-generated findings relate to image evidence, identify unsupported conclusions, detect contradictions or omissions, understand model uncertainty, and compare different AI workflows using reproducible methods. Existing demonstrations rarely provide this level of visibility, making it difficult to evaluate model behavior beyond conventional accuracy metrics. CXR Insight Lab was developed to address this unmet need by shifting the focus from producing another chest X-ray prediction model to creating an auditable multimodal AI workbench. Rather than hiding the internal workflow, the platform exposes every stage—from image preprocessing and visual evidence extraction through structured report generation, deterministic safety validation, and audit logging. The objective is to provide a transparent research environment that supports AI quality assurance, workflow prototyping, education, benchmarking of emerging multimodal foundation models, and responsible evaluation prior to clinical deployment, ultimately promoting greater trust, reproducibility, and accountability in medical imaging AI.

What You Built: Describe your MVP, tool, or workflow in plain language

CXR Insight Lab is a browser-based, end-to-end multimodal AI research workbench designed to make chest X-ray interpretation transparent, explainable, and auditable. Users upload a chest radiograph, after which the platform automatically standardizes the image, associates available annotations, extracts clinically relevant visual evidence using BiomedCLIP, generates a structured radiology report with Qwen3-VL, and evaluates every generated output using a deterministic safety validation framework before presenting results. Unlike conventional AI demonstrations that reveal only the final report, CXR Insight Lab exposes each intermediate processing stage, enabling users to inspect visual evidence, generated observations, structured findings, uncertainty indicators, workflow provenance, and validation outcomes. The platform produces structured JSON reports, visual overlays, audit trails, workflow summaries, and machine-readable exports that support downstream analysis and reproducibility. The deterministic validation engine performs schema validation, evidence consistency assessment, unsupported finding detection, contradiction analysis, completeness checks, and workflow disposition assignment, providing an additional layer of quality assurance before report export. A central design principle of the platform is modularity. Vision-language models, object detection algorithms, segmentation methods, report generation pipelines, and validation policies can all be substituted independently without modifying the overall workflow architecture. This enables rapid benchmarking of emerging multimodal foundation models within a consistent evaluation framework. Rather than functioning solely as a prediction tool, CXR Insight Lab serves as a research infrastructure for AI quality assurance, education, comparative model evaluation, workflow simulation, and transparent development of trustworthy multimodal imaging systems. The prototype is intended exclusively for research and educational purposes and is not designed for clinical diagnosis.

Tech Stack & Development Approach: What tools, frameworks, or methods did you use to build it?

CXR Insight Lab was developed using a modular, research-oriented software architecture emphasizing reproducibility, extensibility, and transparent experimentation. The core workflow is implemented in Python, leveraging OpenCV, NumPy, Pandas, and additional scientific computing libraries for image preprocessing, metadata management, and workflow orchestration. The interactive user interface was built using Gradio to enable rapid deployment of an accessible browser-based research environment without requiring specialized infrastructure. Multimodal visual evidence extraction is performed using BiomedCLIP, which enables image-text representation learning for chest radiographs, while structured radiology reports are generated using the Qwen3-VL vision-language foundation model. Rather than producing free-text outputs alone, the platform represents findings using structured JSON schemas to facilitate downstream validation, interoperability, reproducibility, and automated analysis. A distinguishing component of the platform is its deterministic safety validation engine, which complements generative AI with rule-based verification. The validation framework performs schema validation, evidence consistency analysis, unsupported finding detection, contradiction identification, uncertainty assessment, completeness evaluation, and workflow disposition assignment. Comprehensive audit logging, configurable batch processing, structured exports, and reproducible testing workflows are integrated throughout the pipeline. The application was intentionally engineered as a model-agnostic workbench. Individual vision-language models, segmentation algorithms, object detection methods, reporting pipelines, and validation policies can be exchanged without altering the surrounding infrastructure, enabling standardized benchmarking across evolving multimodal AI systems. Public VinBigData chest radiographs were used for prototype development and evaluation. The complete source code, reproducible Google Colab notebook, documentation, demonstration video, and GitHub repository are publicly available to encourage community evaluation, extension, and collaborative research.

Validation & Evidence: Have you done any informal testing or validation? If so, what did you find?

Validation of the current prototype has focused on demonstrating the feasibility, robustness, and transparency of the complete multimodal workflow rather than establishing diagnostic performance. The platform has been exercised on a 100-image pilot dataset derived from publicly available VinBigData chest radiographs, allowing evaluation of image ingestion, preprocessing, multimodal evidence extraction, structured report generation, deterministic safety validation, audit logging, and export of machine-readable outputs across diverse imaging examples. Beyond successful end-to-end execution, we developed configurable failure injection scenarios designed to challenge the validation framework under controlled conditions. These experiments intentionally introduced unsupported findings, contradictory statements, incomplete reports, schema violations, and inconsistent evidence associations. The deterministic validation engine consistently identified these failures before report export, demonstrating the practical value of combining generative AI with explicit rule-based quality assurance mechanisms. These experiments also highlighted situations where foundation models generated plausible but insufficiently supported findings, reinforcing the importance of transparent validation rather than relying exclusively on model confidence scores. To support reproducibility, the project includes a publicly available GitHub repository, executable Google Colab notebook, configurable test cases, comprehensive documentation, and a video walkthrough illustrating the complete workflow from image upload through audit reporting. Although these results demonstrate technical feasibility and reproducible implementation, the prototype has not undergone formal clinical validation, prospective reader studies, regulatory evaluation, or comparative assessment against practicing radiologists. Consequently, the platform should be viewed as an auditable research workbench for evaluating multimodal AI workflows rather than a clinically validated diagnostic decision-support system.

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

We view CXR Insight Lab as an evolving research platform and are seeking multidisciplinary feedback that will help shape both its technical capabilities and its relevance to future clinical workflows. In particular, we would appreciate input from radiologists, AI researchers, clinical informaticians, educators, software engineers, and healthcare administrators regarding the design of deterministic safety validation policies that balance interpretability, usability, and clinical value. We are especially interested in understanding which forms of evidence visualization, uncertainty communication, workflow provenance, and audit information would be most useful during routine AI-assisted interpretation. We also welcome recommendations for improving human-AI interaction, identifying meaningful workflow quality metrics beyond traditional diagnostic accuracy, and determining how transparent validation frameworks could support responsible AI governance and regulatory readiness. From a research perspective, we hope to collaborate on benchmarking emerging multimodal foundation models, evaluating alternative evidence retrieval strategies, integrating segmentation and detection algorithms, and expanding validation across larger, more diverse imaging datasets. We would also value opportunities to conduct reader studies, educational evaluations, and multi-institutional collaborations that examine how transparency influences user trust, workflow efficiency, and adoption. Finally, we welcome partnerships with clinical imaging groups interested in integrating transparent multimodal AI evaluation into workflow simulation, trainee education, quality assurance initiatives, or prospective research studies. Our broader goal is to build a community-driven platform that enables reproducible comparison of multimodal AI systems while advancing trustworthy, explainable, and auditable imaging workflows that can ultimately support responsible clinical translation.

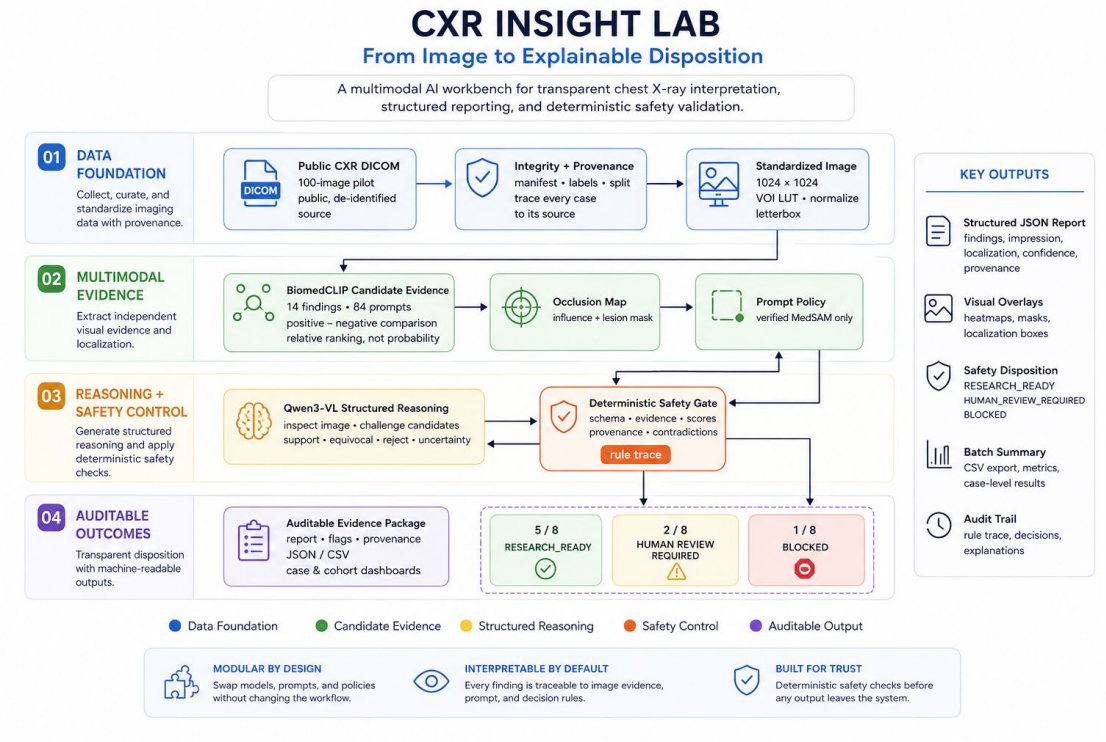
Known Limitations & Honest Failures: What isn't working yet, or what have you already tried that failed?

CXR Insight Lab is intentionally presented as a research prototype rather than a production-ready clinical application. Although the workflow is fully functional, it has not undergone clinical validation, regulatory evaluation, or prospective deployment within healthcare environments. Current testing has been performed primarily on a relatively small pilot dataset of publicly available VinBigData chest radiographs, which cannot fully represent the diversity of patient populations, acquisition protocols, disease prevalence, or reporting complexity encountered in routine clinical practice. Like other multimodal foundation model applications, the quality of generated findings depends on the strengths and limitations of the underlying vision-language models. While the deterministic validation framework successfully identifies many unsupported findings, structural inconsistencies, contradictions, and incomplete reports, it cannot determine absolute clinical correctness or replace expert radiologist interpretation. Transparent validation reduces certain categories of AI failure but does not eliminate hallucinations or guarantee diagnostic accuracy. The current implementation also lacks integration with production imaging infrastructure such as PACS, RIS, Epic, PowerScribe, or DICOM workflow services. Access to vendor APIs, enterprise imaging environments, and larger curated clinical datasets would substantially improve evaluation under realistic operational conditions. Additionally, while exposing intermediate workflow stages increases transparency, it does not fully explain the internal reasoning processes of large multimodal foundation models, highlighting an important limitation shared across contemporary generative AI systems. Future work will emphasize prospective reader studies, large-scale benchmarking, workflow integration, and multi-site validation to establish clinical utility and real-world impact.

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

If fully developed, clinically validated, and integrated into healthcare environments, CXR Insight Lab has the

potential to become a foundational platform for transparent multimodal imaging AI evaluation rather than simply another diagnostic algorithm. By exposing intermediate workflow stages, linking generated findings to supporting visual evidence, and validating outputs through deterministic safety policies, the platform could help establish higher standards for explainability, reproducibility, and quality assurance in medical imaging AI. Radiologists could use the workbench to better understand AI-generated findings, identify unsupported conclusions before report finalization, compare outputs from multiple foundation models, and improve confidence in AI-assisted workflows. AI developers could leverage the modular architecture to benchmark competing vision-language models, detection algorithms, segmentation methods, and validation strategies within a standardized evaluation framework. Educators and trainees could benefit from an interactive learning environment that illustrates how multimodal AI systems process images, generate findings, and respond to failure scenarios, providing valuable insight beyond traditional static demonstrations. Healthcare organizations and clinical informatics teams could employ the platform to prototype imaging workflows, evaluate governance policies, assess model readiness for deployment, and support multidisciplinary AI oversight. More broadly, CXR Insight Lab demonstrates a shift from treating AI as a black-box prediction engine toward viewing it as a transparent, inspectable, and auditable clinical workflow. As multimodal foundation models continue to evolve, platforms that emphasize evidence traceability, deterministic validation, and reproducible benchmarking may play an increasingly important role in accelerating responsible AI adoption, strengthening clinician trust, and improving patient safety across medical imaging.



GitHub Repo:

https://go.screenpal.com/watch/cOiT1xnURED

https://github.com/seemasaharan/cxr-insight-lab

https://github.com/seemasaharan/cxr-insight-lab