



2023-2026

Impact Report

AI Development and Evaluation Lab

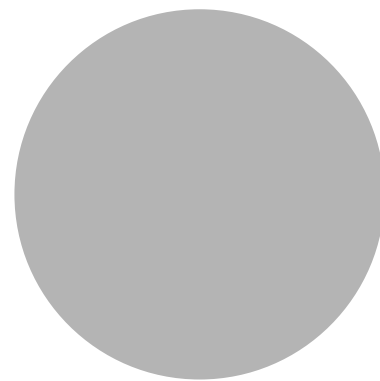
STANFORD UNIVERSITY



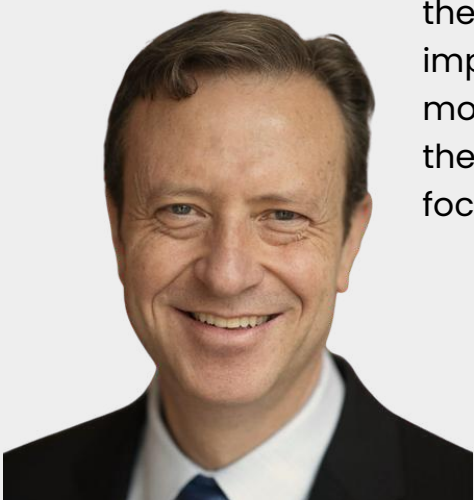
Stanford
MEDICINE

Radiology

Directors' Messages



AI continues to expand what's possible in radiology. But with every new application, development and regulatory clearance are only the first step. Achieving real clinical impact depends on rigorous implementation methods, including systematic evaluation, ongoing monitoring, and targeted feedback. Our lab is focused on building the frameworks and methods that help enable this technology by focusing on the safety, reliability, and transparency of AI.



David Larson, MD, MBA
Founding Director

While AI models often struggle to maintain their initial performance in real-world settings, the mission of the AIDE Lab is to provide the systematic, scalable frameworks necessary to bridge that gap. We ensure that clinicians and patients receive the sustained value promised by these tools. By unifying clinical, technical, IT, and engineering expertise, the AIDE Lab has built a unique environment dedicated to the rigorous oversight of medical AI.

Akshay Chaudhari, PhD
Co-Director



AIDE Lab Overview

Dedicated to systematically improving the safety, reliability, and equity of artificial intelligence in healthcare, primarily focused on medical imaging.

About Us

The AI Development and Evaluation (AIDE) Lab at Stanford Medicine's Department of Radiology advances the responsible development, evaluation, and clinical integration of artificial intelligence in healthcare, with a primary focus on medical imaging. Our mission is to ensure that AI technologies deliver meaningful clinical value while meeting high standards for safety, reliability, and equity, helping bridge the gap between technical innovation and real-world patient care.

The AIDE Lab's work spans the full AI lifecycle, from algorithm development to rigorous evaluation, quality assessment, and post-deployment monitoring. By creating frameworks and tools that assess performance in diverse, real-world clinical settings, the lab supports the adoption of AI systems that clinicians can trust and health systems can sustain. Through research, collaboration, and education, the AIDE Lab contributes to shaping a future where AI is thoughtfully designed and responsibly used to improve healthcare outcomes.

“The work being done by the AIDE Lab reflects the thoughtfulness needed to make sure that only useful AI is deployed, which not only helps radiologists but also reflects good science.”



Andrew Johnston, MD
Pediatric Radiologist
Stanford Radiology

Expanding Our Impact Through Strategic Partnerships

The AIDE Lab is launching a strategic research partnership with Radiology Partners (RP), a leading technology-enabled radiology practice in the U.S, and its technology and AI division, Mosaic Clinical Technologies, to foster the development of artificial intelligence (AI) assessment and continuous monitoring methods and promote the public dissemination of their research. This partnership will leverage RP's extensive experience in implementing AI tools across their national network of radiology practices and draw on the AIDE Lab's expertise in AI assessment and development to create innovative solutions that strengthen the performance, trustworthiness, and accessibility of AI in imaging-based healthcare.

This partnership will generate new insights and establish AI evaluation approaches and tools that can be applied both within Stanford Medicine entities and RP practices, and more broadly, to other institutions. Building on joint research efforts already underway, the partnership will translate RP's real-world learnings from AI deployments into reproducible, peer-reviewed insights. The research will be conducted at Stanford University School of Medicine, Department of Radiology, and will benefit the broader global clinical community by establishing new standards and methods for AI evaluation in radiology.

“We are thrilled to collaborate with AIDE Lab on this important and timely work. This partnership underscores our shared commitment to advancing the responsible development and integration of AI in clinical practice. By working alongside leading academic medical centers like Stanford Medicine, we can accelerate the evaluation and adoption of impactful AI solutions across our national practice and contribute to the broader scientific understanding of how AI can elevate care quality, efficiency and outcomes for patients across the world.”



Nina Kottler, MD
Chief Clinical Artificial Intelligence Officer
Mosaic Clinical Technologies

Clinical Evaluations

The AIDE Lab provides structured, independent evaluations using local data and clinically meaningful metrics to determine if models are fit for their intended purpose, reliable across patient groups, and aligned with clinical needs. This ensures that only the AI tools with demonstrated utility and validated performance are implemented, enabling radiologists and clinics to effectively adopt AI.

Clinical Evaluations Overview

9 Evaluations Completed

6 Evaluations in progress

3 Models passed evaluation

Moved to implementation with Stanford Technology and Digital Solutions (TDS)

2 Models failed evaluation

Pre-deployment evaluations saved up to:

- **1384** hours of TDS labor
- **\$47,000** of server and storage costs
- **\$240,000** of annual vendor fees

4 models deprioritized

Built for internal and vendor evaluations

The AIDE Lab's AI evaluation process allows for performance assessments of both internal and vendor models. To date, the evaluations have spanned across:

15
Vendors

6
Internal
Tools



Being innovative in AI isn't about having the most models implemented, it's being able to curate models and identify the ones with value.



Gloria Hwang, MD

Interim Vice Chair of Clinical Informatics
Stanford Radiology



Research Highlights

The AIDE Lab conducts research focused on advancing the development, evaluation, and monitoring of artificial intelligence in medical imaging, with an emphasis on real-world clinical impact. Its work spans algorithm development, methods for rigorous and equitable evaluation, and strategies for assessing quality and performance over time, including after deployment in clinical settings. By studying how AI behaves across diverse patient populations, imaging conditions, and workflows, the lab generates evidence-based frameworks and tools that help ensure AI systems are reliable, trustworthy, and clinically meaningful.

Project Spotlights

Assessing completeness of clinical histories accompanying imaging orders using adapted open-source and closed-source large language models

- Both open-source and closed-source large language models (LLMs) were adapted to automatically extract key elements of clinical histories from radiology imaging orders
- A fine-tuned open-source model (Mistral-7B) performed comparably to a proprietary model (GPT-4 Turbo), achieving high accuracy and substantial agreement with radiologist annotations
- When applied to a large dataset of approximately 49,000 clinical histories, only about one-quarter contained all five key elements, indicating that many imaging orders lack complete clinical information
- The findings suggest that LLM-based tools could be used to benchmark and monitor the completeness and quality of clinical histories, potentially supporting systematic quality improvement efforts in radiology

Input

See comments Order History: Relevant PMH CAD s/p PCI, HTN, TIIDM. Presents with chest pain, marked HTN for a duration of 1 day. Specific location of issue (if applicable): . Concern for dissection.

Output

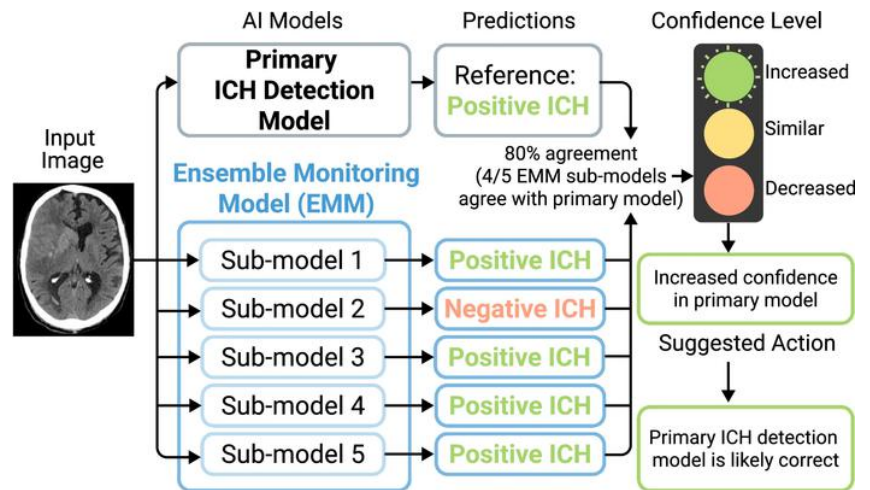
Element	Fine-tuned Mistral	Reader 1	Reader 2
Past Medical History	coronary artery disease status post percutaneous coronary intervention, hypertension, type II diabetes mellitus	Coronary artery disease status post percutaneous intervention, hypertension, type II diabetes mellitus	coronary artery disease status post percutaneous coronary intervention, hypertension, diabetes mellitus type 2
What	chest pain, marked hypertension	Chest pain, marked hypertension	chest pain, marked hypertension
When	1 day	1 day	1 day
Where	chest	Chest	chest
Clinical Concern	dissection	Dissection	dissection

Figure. Example input clinical history entry with many clinical abbreviations that expert radiologists would be expected to recognize and that the fine-tuned Mistral model also correctly annotated. Reader 1 and 2 indicate two radiologists who manually annotated the clinical history.

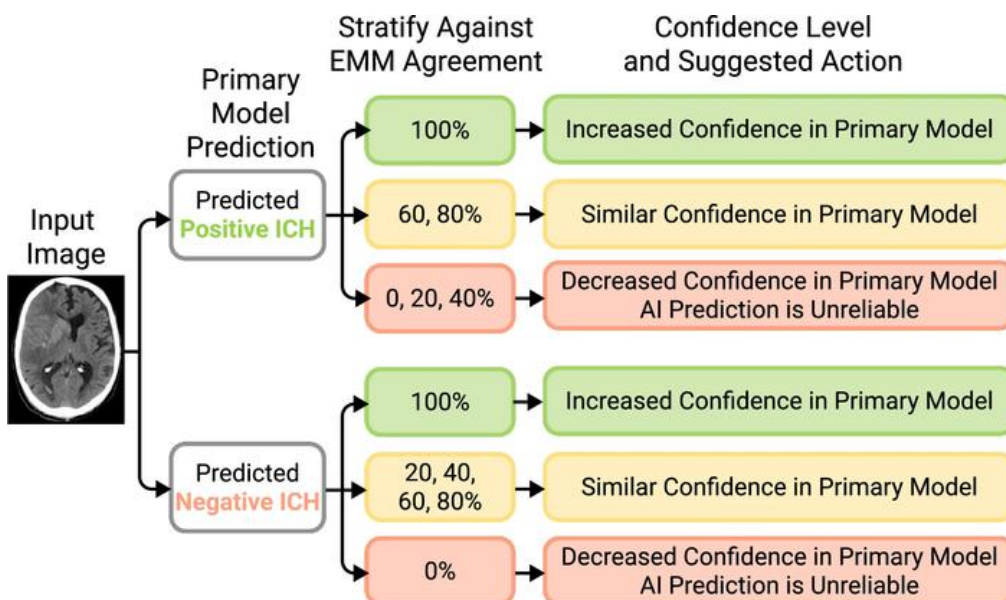
Project Spotlights

Automated real-time assessment of intracranial hemorrhage detection AI using an ensembled monitoring model (EMM)

- Current AI tools in radiology lack real-time performance monitoring once deployed, which can increase cognitive burden on clinicians and potentially reduce trust in AI predictions because users must independently gauge prediction reliability
- The Ensembled Monitoring Model (EMM) that operates alongside commercial “black-box” AI tools to estimate confidence for individual AI predictions in real time, without needing access to internal model details
- When applied to intracranial hemorrhage detection on head CT scans, EMM can stratify predictions by confidence level, helping identify cases where the primary model’s output was likely reliable versus those warranting closer radiologist review, potentially reducing cognitive workload and improving decision support
- The EMM approach shows promise as a deployable monitoring solution for real-world clinical AI systems by providing actionable confidence estimates that can inform how clinicians use AI results at the point of care, supporting safer and more trustworthy integration of AI into radiology practice

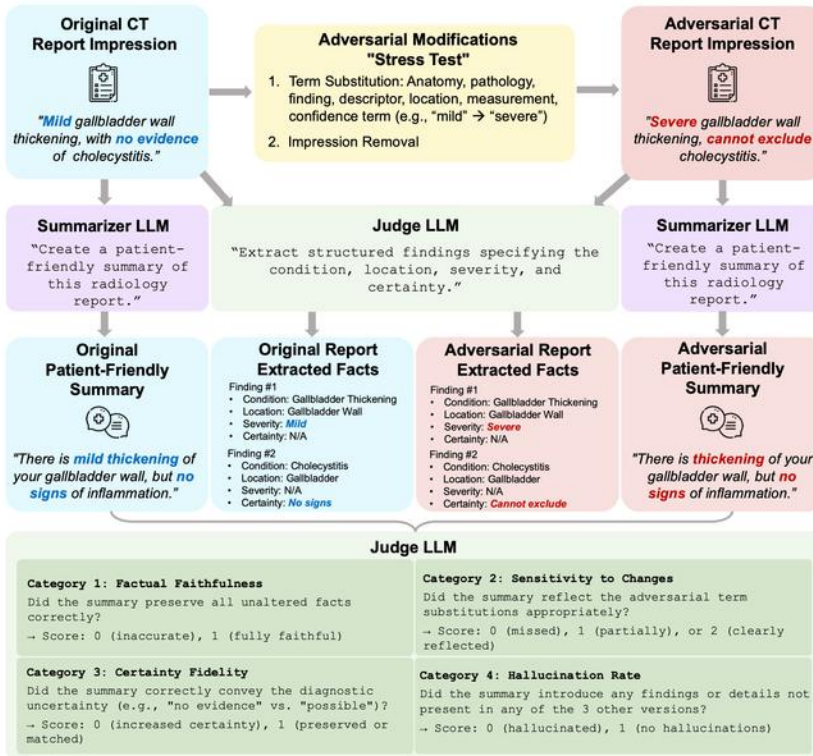


The EMM approach: Each sub-model within the EMM is trained to perform the same task as the primary ICH detection model. The independent sub-model outputs are then used to compute the level of agreement between the primary ICH detection model and EMM, helping quantify confidence in the reference prediction and suggesting an appropriate subsequent action.



An example EMM use case is stratifying cases into categories of increased, similar, or decreased confidence in the primary model predictions based on the level of EMM agreement

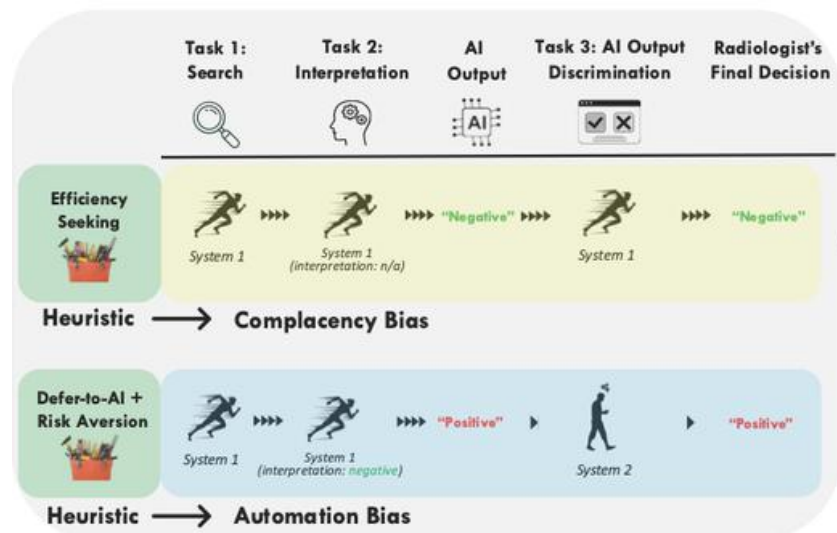
Can AI be trusted to summarize radiology reports for patients? A robustness evaluation using adversarial edits



- Large language models can generate patient-friendly summaries of radiology reports, but adversarial edits reveal consistent vulnerabilities, particularly around missed descriptors, confidence terms, and subtle wording changes that affect clinical meaning.
- A scalable, automated "stress-testing" framework can systematically identify failure modes, such as reduced sensitivity to term substitutions and occasional hallucinations. This highlights risks that may not appear in standard evaluations.
- Model performance varies across architectures, with some models demonstrating stronger factual faithfulness and uncertainty preservation, underscoring the need for rigorous evaluation before deploying patient-facing AI tools.

A framework for understanding cognitive biases in AI-assisted diagnostic radiology

- Effective use of AI in diagnostic radiology depends not only on model performance, but on how radiologists cognitively interact with AI outputs.
- In addition to the two main cognitive tasks of searching and interpretation that radiologists must perform for each case, the use of AI introduces a novel third cognitive task: AI output discrimination.
- Applying principles from cognitive sciences and human factors research, a new framework delves into the heuristics at play and the possible cognitive biases that can arise, allowing a deeper understanding of how radiologists use AI tools - beyond simply "over-reliance" or "under-reliance" - and enabling tailored mitigation strategies.



Research Overview



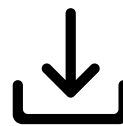
5

Publications



63

Citations



15K+

Reads

Assessing Completeness of Clinical Histories Accompanying Imaging Orders Using Adapted Open-Source and Closed-Source Large Language Models. Radiology. 2025 February;314(2). [doi: 10.1148/radiol.241051](https://doi.org/10.1148/radiol.241051)

Automated real-time assessment of intracranial hemorrhage detection AI using an ensembled monitoring model (EMM). Npj Digital Medicine, 8(1). [doi: s41746-025-02007-0](https://doi.org/10.1038/s41746-025-02007-0)

Foundation Models in Radiology: What, How, Why, and Why Not. Radiology, 2025 February;314(2). [doi: 10.1148/radiol.240597](https://doi.org/10.1148/radiol.240597)

Predicting the Value of Radiology Artificial Intelligence Applications: Large-Scale Predeployment Evaluation of a Portfolio of Models. AJR, 2026 March. [doi: 10.2214/AJR.25.34340](https://doi.org/10.2214/AJR.25.34340)

Design for Understanding: Sentence-Level Cross-Referencing Improves Comprehension and Confidence in AI-Generated Patient-Friendly Radiology Reports. JACR, 2026 March. [doi: 10.1016/j.jacr.2026.02.037](https://doi.org/10.1016/j.jacr.2026.02.037)

Conference Presentations



Awards

2 Magna Cum Laude Awards at RSNA 2025

1 2024 ACR QS+I Innovation Award

1 Human-Centered Artificial Intelligence Seed Grant

Coverage in the Media



AI Can Help Determine the Completeness of Clinical Histories

[View here](#)



Foundation models are medtech's newest trend. But what are they?

[View here](#)



Foundation Models in Radiology: The Future of Medical Imaging

[View here](#)



AuntMinnie.com

Stanford AIDE lab proposes radiology AI safety monitoring model

[View here](#)



Monitoring AI with AI

[View here](#)



Predeployment Evaluation of the Value of Radiology AI Models

[View here](#)



Education and Outreach

During the summer of 2025, the AIDE Lab hosted a summer internship program designed for high school and undergraduate students passionate about AI, machine learning, data science, and the clinical application of AI tools. This program offered hands-on experience in AI research, model development, performance evaluation, and monitoring, providing an opportunity to work alongside experienced researchers. Interns gained practical knowledge, contributed to innovative projects, and enhanced their technical and problem-solving skills.



**NATHAN
LE**

Senior
UCLA



**ALARA
MARTIN**

Senior
Palo Alto High School



**SAMHITA
LAXMAN**

Senior
Gunn High School

Interest in the summer program was immense, with over 130 applications received for only a few spots. The selected summer interns pursued complementary projects that advanced practical AI research in medical imaging and model performance. Undergraduate researcher Nathan Le focused on improving the accuracy of existing vendor models through test-time augmentations, a technique that enhances model performance without retraining, while also engaging with the lab's interdisciplinary work to inform his future studies at the intersection of healthcare and AI. Meanwhile, high school students Samhita Laxman and Alara Martin collaborated on releasing an annotated dataset and corresponding machine learning model to evaluate the quality of pediatric pelvic radiographs, as well as a paper that explores and assesses pelvic positioning. Through these projects, they gained hands-on experience with real clinical datasets and AI implementation.

Our Team

Current Members

David Larson, MD, MBA

Founding Co-Director

Akshay Chaudhari, PhD

Co-Director

Hye Sun Na

Managing Director

Arogya Koirala, MS

Machine Learning Engineer

Bonnie Armstrong, PhD

Postdoctoral Scholar

Magdalini Paschali, PhD

Postdoctoral Scholar



Collaborators

Shannon Walters, MS, ARRT(R)(MR)

Executive Technical Director

3D and Quantitative Imaging (3DQ) Lab

Derrick Laurel, RT(R)

Clinical AI Lead Technologist

3D and Quantitative Imaging (3DQ) Lab

Lina Cheuy, PhD

Technical Writer & Communications

Manager, Stanford Radiology

Alumni

Zhongnan Fang, PhD

Former Principal Machine

Learning Scientist

Camila Gonzalez, PhD

Former Postdoctoral Scholar

Lab Sponsors

Garry Gold, MD, MS

Professor and Former Chair

Stanford Radiology

Yun-Ting Yeh, MBA

Director, Finance & Administration

Stanford Radiology