Scientific Journals
An international open-access platform

Hanlin Letters

Article

OPEN ACCESS

Clinician Artifact Literacy Training for Safer Use of Medical Vision-Language Models During Diagnostic Imaging Education Sessions

Ahmed Fawzy,[1] Omar Khaled,[2] Tarek Mostafa[3]

[1] Faculty of Computers and Information, South Valley University, Qena 83523, Egypt;
[2] Department of Management Information Systems, Sadat Academy for Management Sciences, 54 Corniche El Nile Street, Maadi, Cairo 11431, Egypt;
[3] Faculty of Commerce and Business Administration, Helwan University, Helwan District, Cairo 11795, Egypt;

Abstract

Medical vision-language models are increasingly shown to students, residents, and clinicians as tools for image explanation, case discussion, and preliminary educational interpretation. Their outputs can be persuasive even when the image contains acquisition artifacts, overlays, degraded contrast, or nonclinical visual marks. This creates an educational risk: learners may absorb the model's explanation without recognizing that the answer has been shaped by image conditions rather than by reliable diagnostic evidence. This paper reports a controlled educational study of artifact literacy training for medical vision-language model use in diagnostic imaging education. We enrolled 186 participants across medical students, radiology residents, emergency medicine residents, and physician assistants. Participants reviewed model-assisted imaging cases before and after a structured training module focused on recognizing artifact-sensitive model behavior, distinguishing image evidence from model narrative, and documenting uncertainty. The intervention used chest radiography, musculoskeletal radiography, dermatology photographs, and ultrasound frames with benign artifacts, clinically relevant artifacts, and model-salient nonclinical marks. Compared with the control group, trained participants improved their artifact-related error detection from 41.8% to 73.6%, reduced inappropriate acceptance of model explanations by 28.9 percentage points, and wrote more specific uncertainty notes. The largest gains appeared among medical students and non-radiology trainees. Follow-up testing four weeks later showed partial retention, with artifact-related error detection remaining 19.4 percentage points above baseline. The findings suggest that safe educational use of medical vision-language systems requires explicit training in artifact literacy rather than simple warnings about model limitations.

1. Introduction

Medical learners increasingly encounter artificial intelligence outputs while studying imaging cases. A vision-language model may describe a chest radiograph, answer a dermatology question, explain a fracture pattern, or summarize an ultrasound frame in natural language. In an educational setting, this is attractive because the model can produce immediate feedback and can phrase image findings in accessible terms. A learner can ask what is visible, why a finding matters, and how a visual sign relates to a diagnosis. The same interaction, however, can also create a new form of instructional error. The model's answer may sound coherent even when it has been influenced by an artifact, a marker, a visual overlay, poor acquisition quality, or a misleading image feature.

This paper studies that instructional error from the perspective of learner preparation. The focus is not on building a stronger model. It is not on creating a new diagnostic benchmark. It is not on routing user intent, auditing provenance, federated training, reproducibility logging, or mathematical risk estimation. The focus is a practical educational question: can clinicians and trainees be taught to recognize when a medical vision-language model's answer may have been distorted by image artifacts?

Artifact literacy is the ability to notice visual conditions that can influence image interpretation and model behavior. In human radiology training, learners are taught about motion blur, portable technique, rotation, exposure,

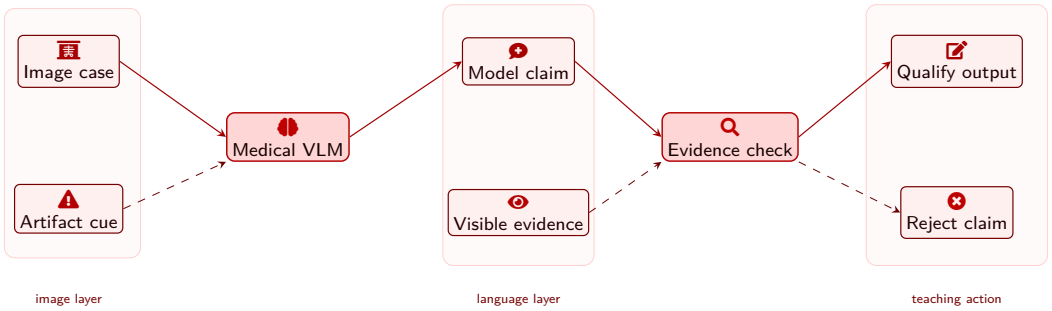

Figure 1: Artifact-sensitive model review pipeline linking image conditions, generated explanation, evidence checking, and calibrated instructional response.

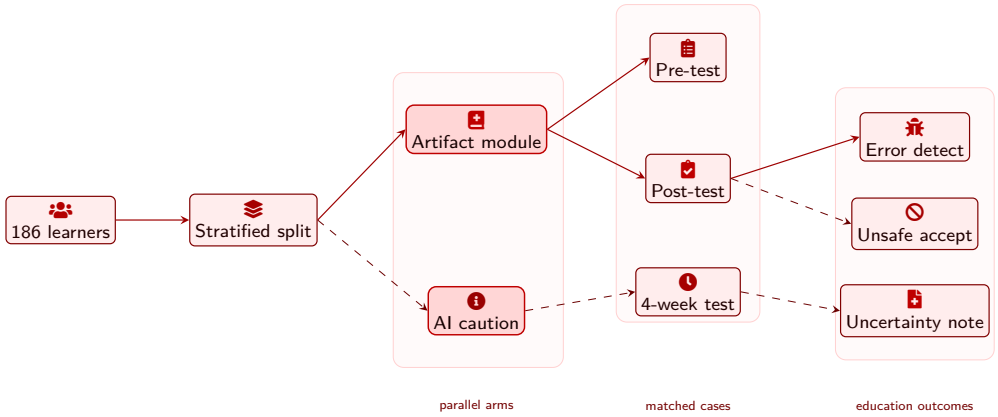

Figure 2: Controlled educational study structure comparing artifact-focused instruction against a general AI caution module across repeated assessments.

skin folds, external objects, surgical material, and acquisition limitations. In model-assisted learning, artifact literacy needs an additional layer. The learner must recognize not only how an artifact affects human interpretation, but also how it may affect the model's language. A model may overread a grid pattern, mistake an external marker for pathology, treat a ruler or arrow as meaningful evidence, or become overly confident after seeing a nonclinical cue. It may also ignore an artifact and produce a standard explanation that does not acknowledge image limitations.

A warning that "AI can be wrong" is too general to teach this skill. Learners need to practice with concrete cases. They need to see how a model changes its answer when the same image is shown with a subtle overlay, when a symbol is placed near a lesion, when contrast is degraded, or when an external object appears in the field. They need to learn a response habit: inspect the image first, identify possible artifacts, read the model answer critically, compare the model's stated evidence with visible findings, and write uncertainty when the evidence is limited. The present study evaluates whether a brief structured training module can produce that habit.

The study was conducted in an educational simulation rather than a clinical deployment. Participants were not asked to make patient-care decisions. They reviewed de-identified images and model responses, judged whether the model's explanation should be accepted, edited, or rejected, and wrote short comments explaining their decision. This design allowed measurement of educational outcomes while avoiding clinical risk. The aim was to test whether artifact literacy can be taught and retained, not whether model-assisted diagnosis should be used in practice.

The educational problem is becoming more important because model demonstrations often emphasize impressive cases. Learners may see examples where a model correctly identifies an abnormality and explains it fluently. They may see fewer examples where the model is misled by an acquisition artifact or nonclinical mark. This can create an unrealistic sense of reliability [1]. A balanced curriculum should include ordinary correct cases, ordinary

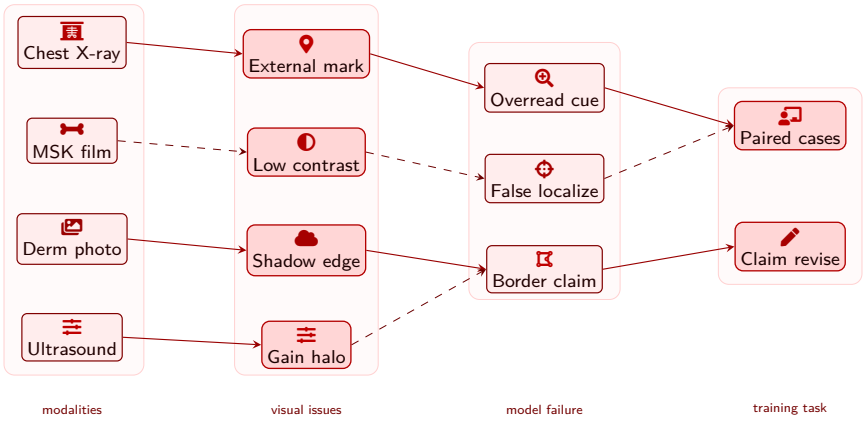

Figure 3: Cross-modality artifact map showing how acquisition and nonclinical visual conditions can shape model language and case-based teaching tasks.

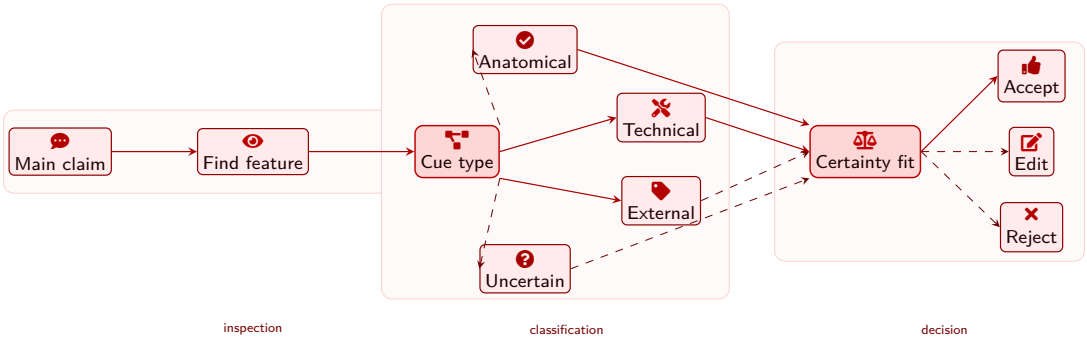

Figure 4: Compact review checklist for judging whether a model explanation is supported by anatomical evidence, technical context, or uncertain visual cues.

incorrect cases, and cases where the source of error is the interaction between the model and the visual condition of the image.

The present study used a randomized controlled design. Participants were assigned to either artifact literacy training or a general AI caution module. Both groups learned that medical AI systems can make mistakes. Only the intervention group received structured practice in identifying artifact-sensitive model outputs. Both groups completed the same pre-test, immediate post-test, and four-week follow-up test. The primary outcome was artifact-related model error detection. Secondary outcomes included inappropriate acceptance of model answers, quality of uncertainty documentation, time per case, and participant confidence calibration.

The training module was built around four educational principles. First, learners should distinguish image quality issues from diagnostic findings. Second, learners should identify when a model's explanation refers to a visual feature that may be artificial or external. Third, learners should avoid accepting model certainty when the image evidence is weak. Fourth, learners should document what remains uncertain rather than simply saying the model is wrong. These principles were presented through short examples, guided practice, and feedback.

Several medical VLM studies have used visual perturbations to test model behavior, but those perturbations are usually framed as technical evaluation tools. Here, they are reframed as educational materials. In one relevant methodological detail, Sadanandan and Behzadan (2025) described visual perturbation methods including Gaussian noise, checkerboard overlays, random arrows, Moiré patterns, and steganographic least-significant-bit techniques for probing model robustness [2].

The paper contributes an educational dataset, a training intervention, and empirical evidence about learner outcomes. It also proposes a practical assessment rubric for artifact literacy. The results show that a short training module improves detection of artifact-related model errors and reduces inappropriate acceptance of model-generated

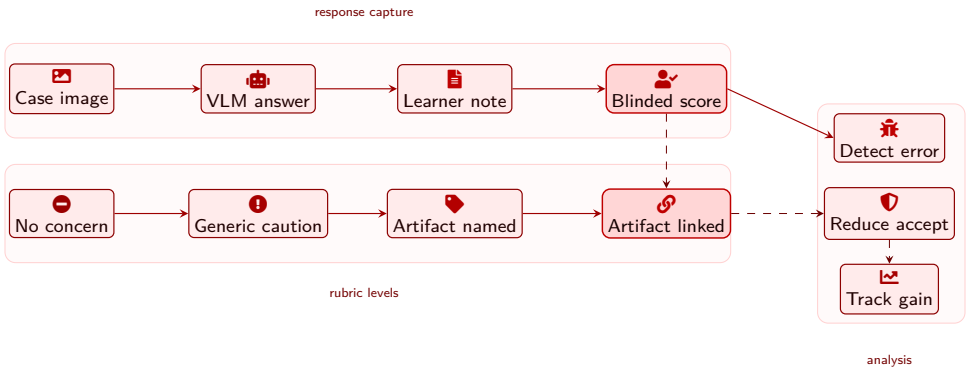

Figure 5: Assessment rubric pathway converting learner comments into scored artifact-literacy evidence and outcome measures for instructional evaluation.

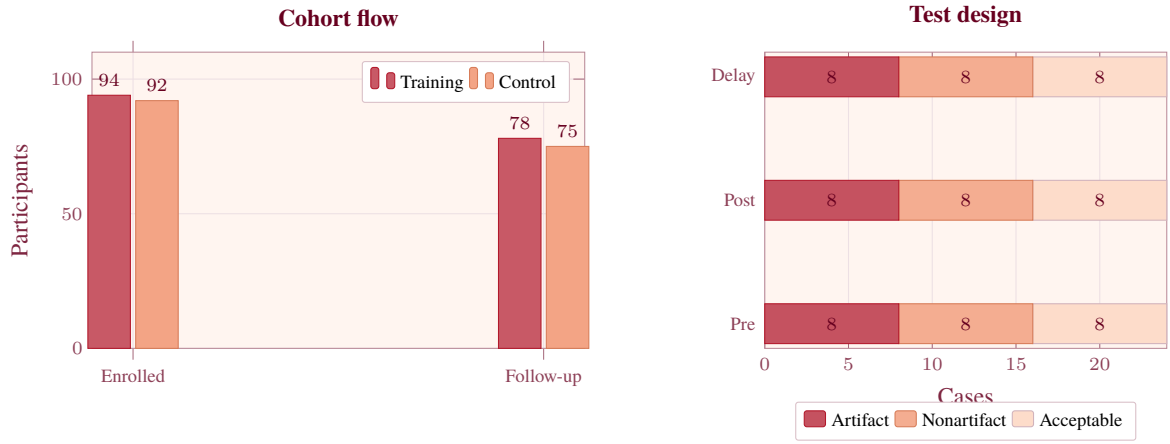

Figure 6: Randomized study structure showing enrollment, four-week follow-up counts, and balanced 24-case assessment composition across the pre-test, immediate post-test, and delayed follow-up.

explanations. The gains are not uniform across learner groups, and retention declines over time, but the results support artifact literacy as a teachable competency.

2. Study Setting and Participants

The study was conducted across two academic medical centers and one regional teaching hospital. Participants were recruited from scheduled educational sessions and voluntary professional development workshops. The final sample included 186 participants. There were 64 medical students, 43 radiology residents, 39 emergency medicine residents, 24 physician assistants, and 16 attending clinicians who participated as continuing education learners. Participants were eligible if they had prior exposure to medical imaging education but had not completed a formal course on vision-language model limitations [3].

The study used a randomized parallel-group design. Participants were assigned to either the artifact literacy intervention or the general AI caution control. Randomization was stratified by training level so that the groups had similar distributions of students, residents, and practicing clinicians. The intervention group contained 94 participants, and the control group contained 92 participants [4]. The groups were similar in prior AI exposure, self-rated imaging confidence, and average years of clinical training.

All participants completed a pre-test before receiving any instructional module. They then completed either the intervention or the control module. Immediately after the module, they completed a post-test with new cases. Four weeks later, they completed a follow-up test online. The follow-up response rate was 82.3%, with 78 intervention participants and 75 control participants completing the delayed assessment.

Table 1: Participant Distribution by Training Level

Group	Participants	Baseline Detection (%)	Post-Training (%)
Medical Students	64	31.4	68.2
Radiology Residents	43	58.9	81.4
EM Residents	39	39.7	71.5
Physician Assistants	24	36.9	66.7
Attending Clinicians	16	55.6	76.2

Table 2: Primary Outcome Comparison Between Groups

Metric	Intervention	Control	Difference
Pre-Test Detection (%)	41.8	43.4	–1.6
Post-Test Detection (%)	73.6	51.2	22.4
Improvement (%)	31.8	7.8	23.9
Follow-Up Detection (%)	61.2	48.3	12.9

The educational cases were de-identified and selected from four imaging categories. Chest radiography cases included portable radiographs, external markers, tubes, and common cardiopulmonary findings. Musculoskeletal radiography cases included fractures, orthopedic hardware, positioning issues, and external objects. Dermatology cases included lesion photographs with lighting variation, ruler markings, shadow, scale, and color imbalance. Ultrasound frames included shadowing, gain variation, probe-pressure effects, and annotation overlays.

Each case included an image, a short question, and a model-generated answer. The model answer was sometimes correct, sometimes incorrect for ordinary reasons, and sometimes incorrect because it appeared to rely on an artifact or nonclinical visual cue. Participants were asked to choose one of three decisions: accept the model answer, accept with edits or uncertainty, or reject the model answer. They also wrote a brief explanation.

The cases were divided into three assessment sets: pre-test, immediate post-test, and delayed follow-up. No image appeared in more than one assessment set. The assessment sets were matched by modality, difficulty, and error type. Each test contained 24 cases. Eight cases contained artifact-related model errors, eight contained non-artifact model errors, and eight contained acceptable model answers. This structure allowed the study to measure whether training specifically improved artifact-related detection rather than simply increasing skepticism toward all model outputs.

The intervention module lasted 45 minutes. It was delivered as a guided workshop in person or through a live video session. The control module also lasted 45 minutes and covered general limitations of AI in medicine, including bias, hallucination, privacy, and the need for human oversight. The control group did not receive case-based artifact training. This active control was used to distinguish the effect of artifact-specific practice from general caution about AI [5].

The study was reviewed as an educational research protocol. Images were de-identified, no patient-care decisions were made, and participant data were analyzed without personal identifiers. Participants gave consent for their assessment responses to be used in aggregate educational research. They were told that the study evaluated instructional methods, but they were not told the specific hypothesis until after completing the final assessment [6].

3. Training Intervention

The artifact literacy module was designed as a compact educational intervention that could fit within a residency conference, clerkship teaching session, or continuing education workshop. It did not require programming knowledge. It did not teach model architecture. It taught a practical review routine for model-assisted image interpretation [7].

The module had four parts.

- The first part introduced common visual conditions that can influence both human readers and models. Examples included low contrast, rotation, motion blur, external objects, overlying lines, skin folds, ruler markings, compression artifacts, and annotation symbols.
- The second part showed paired model responses. Participants saw a clean image response and an artifact-altered image response. They discussed which words in the model output appeared to be driven by the artifact [8].
- The third part taught an evidence-checking routine. Participants practiced asking whether the model's

Table 3: Artifact-Related Acceptance Patterns

Condition	Pre-Test (%)	Post-Test (%)	Change
Intervention Acceptance	46.7	17.8	–28.9
Control Acceptance	44.9	36.1	–8.8
High-Confidence Errors (Intervention)	29.4	8.7	–20.7
High-Confidence Errors (Control)	28.1	21.9	–6.2

Table 4: Written Explanation Quality Scores

Category	Intervention Pre (%)	Intervention Post (%)	Control Post (%)
Specific Artifact Linked to Claim	22.5	61.3	32.4
Generic Concern Only	41.8	18.7	37.2
Artifact Mentioned Without Link	23.1	14.6	19.5
No Meaningful Concern	12.6	5.4	10.9

stated reason was visible, whether the visible feature was anatomical or external, whether the finding could be explained by acquisition technique, and whether the model's certainty exceeded the image evidence.

- The fourth part used guided correction. Participants rewrote model answers to include appropriate uncertainty or to remove unsupported claims.

The module emphasized concise habits rather than long theoretical explanations. Participants were encouraged to use a four-step verbal checklist while reviewing model output:

- identify the model's main claim;
- locate the image feature used to support the claim;
- decide whether that feature is anatomical, technical, external, or uncertain;
- revise the answer if the support is weak or artifact-sensitive.

The control module had the same length and similar presentation style. It described broad concerns about medical AI, including general error, bias, privacy, overreliance, and lack of accountability. It included examples of incorrect model outputs, but it did not teach artifact categories, visual cue checking, or model-answer editing. The control was intentionally credible so that any difference between groups would not simply reflect receiving more attention or more warnings.

The intervention used examples that were realistic but not clinically high stakes. For example, a chest radiograph case showed a model interpreting an external line as a support device. A musculoskeletal case showed a model overemphasizing a skin marker near a suspected fracture. A dermatology case showed a model describing border irregularity after being influenced by shadow and ruler placement. An ultrasound case showed a model overstating a lesion boundary when the frame included gain-related edge enhancement.

Participants were not told to distrust all model outputs. The curriculum stressed that model answers can be useful when they are tied to visible evidence and appropriately qualified. The goal was calibrated use rather than blanket skepticism. This distinction was important because over-skepticism can also harm educational value. A learner who rejects every model response has not learned artifact literacy; they have only learned avoidance [9].

The intervention also included examples where the artifact was present but the model answer remained acceptable. This prevented participants from treating every artifact as a reason to reject the model. A portable chest radiograph may have low quality but still show a clearly visible pneumothorax. A dermatology photograph may contain a ruler but still support a description of color variation. The task is not to reject images with artifacts; the task is to judge whether the model's claim depends on the artifact or exceeds the evidence.

4. Assessment Measures

The primary outcome was artifact-related model error detection. A participant received credit when they correctly identified that a model answer should not be accepted because the answer was influenced by or failed to account for an artifact. This outcome was measured as the percentage of artifact-related error cases correctly marked as reject or accept with edits. Responses that accepted the model answer without qualification were counted as missed detections.

Secondary outcomes included inappropriate acceptance, uncertainty documentation quality, over-rejection, time per case, and confidence calibration. Inappropriate acceptance was the percentage of artifact-related error

Table 5: Performance Across Imaging Modalities

Modality	Baseline (%)	Post-Test (%)	Improvement
Chest Radiography	49.1	77.3	28.2
Musculoskeletal Imaging	45.4	74.6	29.2
Dermatology Images	35.2	71.8	36.6
Ultrasound Frames	37.5	70.6	33.1

Table 6: Assessment Structure and Case Composition

Assessment Component	Cases	Purpose	Distribution
Artifact-Related Errors	8	Error Detection	Balanced
Non-Artifact Errors	8	General Critique	Balanced
Acceptable Outputs	8	Over-Rejection Check	Balanced
Total Per Assessment	24	Full Evaluation	Equal Split

cases accepted without edits. Uncertainty documentation quality was scored on a four-level rubric [10]. A high-quality note identified the specific visual issue and stated how it limited the model answer. A low-quality note used vague language such as "AI may be wrong" without tying the concern to the image.

Over-rejection was measured on acceptable model-answer cases. If a participant rejected a correct and adequately supported model answer, that response counted as over-rejection. This measure was included because the intervention could have made participants overly suspicious. A useful training module should reduce unsafe acceptance without causing broad rejection of correct answers.

Confidence calibration was measured by asking participants to rate confidence in each decision on a five-point scale. Calibration was assessed by comparing confidence with correctness. The study did not expect perfect calibration, but it examined whether training reduced high-confidence acceptance of artifact-related errors.

Time per case was recorded automatically in the online assessment platform. For in-person sessions, participants completed cases through the same platform on tablets or laptops. The time measure was exploratory. A training module that improves performance only by greatly increasing review time may be difficult to implement in real educational settings.

The written explanations were scored by two blinded reviewers. Reviewers did not know participant group or assessment time. They used a rubric with four categories:

- no meaningful concern stated;
- generic concern about AI or image quality;
- specific artifact identified but not connected to the model answer;
- specific artifact identified and connected to the model answer or uncertainty.

Disagreements were resolved by discussion. Reviewer agreement for the explanation rubric was strong, with weighted agreement of 0.82. For the primary detection outcome, automated scoring from participant choices was used, with manual review only for ambiguous written comments.

Statistical analysis compared change from pre-test to post-test between groups. The main comparison was the difference in improvement for artifact-related error detection. Secondary analyses examined subgroup effects by training level and modality. The delayed follow-up measured retention. A result was treated as educationally meaningful if the intervention improved artifact-related detection without increasing over-rejection by more than 5 percentage points.

5. Results

At baseline, participants detected artifact-related model errors in 42.6% of cases. There was variation by training level. Radiology residents had the highest baseline detection at 58.9%, while medical students had the lowest at 31.4%. Emergency medicine residents and physician assistants fell between those groups. Baseline detection was highest for obvious external objects and lowest for subtle lighting, ultrasound gain, and mild radiographic positioning artifacts.

After training, the intervention group improved from 41.8% to 73.6% artifact-related error detection. The control group improved from 43.4% to 51.2%. The between-group difference in improvement was 23.9 percentage points. The intervention effect was consistent across modalities, although the largest gains occurred in dermatology and ultrasound cases. These were the domains where baseline artifact recognition was weakest.

Table 7: Core Components of the Training Module

Module Component	Educational Focus	Activity Type
Artifact Introduction	Visual condition recognition	Guided examples
Paired Comparisons	Output sensitivity analysis	Side-by-side review
Evidence Checking	Claim verification routine	Structured checklist
Guided Correction	Revising unsupported claims	Editing exercises

Table 8: Review Checklist Used During Training

Checklist Step	Purpose	Target Skill
Identify Main Claim	Define model conclusion	Comprehension
Locate Supporting Feature	Match text to image evidence	Evidence tracing
Classify Feature Type	Distinguish artifact vs anatomy	Artifact literacy
Revise if Needed	Add edits or uncertainty	Safe interpretation

Inappropriate acceptance of artifact-related model answers decreased sharply in the intervention group. Before training, intervention participants accepted artifact-influenced model answers without edits in 46.7% of such cases. After training, this fell to 17.8%. In the control group, inappropriate acceptance decreased from 44.9% to 36.1%. The intervention therefore reduced inappropriate acceptance by 28.9 percentage points, compared with 8.8 percentage points in the control group.

The intervention did not produce broad over-rejection [11]. On acceptable model-answer cases, over-rejection increased from 9.6% to 12.8% in the intervention group. In the control group, over-rejection increased from 9.1% to 10.7%. The between-group difference was small and below the predefined educational concern threshold. This suggests that participants became more selective rather than simply more distrustful.

Written explanation quality improved substantially. Before training, only 22.5% of intervention-group comments identified a specific artifact and connected it to the model answer. After training, this rose to 61.3%. Control participants improved from 23.8% to 32.4%. The most common improvement was moving from a generic warning such as "AI could be wrong" to a specific note such as "the bright external marker may be influencing the model's localization claim" or "the shadow at the lesion edge makes the border description uncertain."

Confidence calibration improved. Before training, participants were often highly confident when accepting artifact-related model errors. In the intervention group, high-confidence inappropriate acceptance decreased from 29.4% to 8.7%. In the control group, it decreased from 28.1% to 21.9%. The intervention therefore reduced the most concerning pattern: confidently trusting a model answer shaped by a visual artifact.

Time per case increased modestly. Intervention participants spent an average of 64 seconds per case at baseline and 78 seconds after training. Control participants increased from 63 seconds to 69 seconds. The additional time in the intervention group was concentrated in artifact-related cases. For acceptable model answers, review time changed little. This suggests that trained participants did not slow down uniformly; they spent more time when the case contained a reason for caution.

Subgroup analysis showed the largest improvement among medical students. Their artifact-related detection increased from 31.4% to 68.2%. Emergency medicine residents improved from 39.7% to 71.5%. Physician assistants improved from 36.9% to 66.7%. Radiology residents improved from 58.9% to 81.4%. Attending clinicians improved from 55.6% to 76.2%, although this subgroup was small. The findings suggest that radiology training provides some baseline artifact literacy, but explicit model-focused training still adds value.

Modality-specific outcomes showed that chest radiography improved from 49.1% to 77.3% detection in the intervention group. Musculoskeletal radiography improved from 45.4% to 74.6%. Dermatology improved from 35.2% to 71.8%. Ultrasound improved from 37.5% to 70.6%. The control group showed smaller improvements in every modality. Dermatology and ultrasound remained somewhat harder after training, but the gap narrowed.

The four-week follow-up showed partial retention. Intervention participants detected artifact-related model errors in 61.2% of cases at follow-up, compared with 41.8% at baseline and 73.6% immediately after training. Control participants detected 48.3% at follow-up, compared with 43.4% at baseline and 51.2% immediately after the control module. The intervention group therefore

Table 9: Retention Outcomes After Four Weeks

Measure	Baseline (%)	Immediate Post (%)	Follow-Up (%)
Artifact Detection	41.8	73.6	61.2
Specific Explanations	22.5	61.3	47.9
Over-Rejection	9.6	12.8	11.9
Confidence Calibration Gain	0.0	20.7	13.5

Table 10: Educational Recommendations for Artifact Literacy

Recommendation	Purpose	Expected Benefit
Include Clean Cases	Prevent excessive skepticism	Balanced judgment
Use Artifact-Sensitive Cases	Teach visual critique	Safer review habits
Require Response Editing	Encourage nuanced decisions	Better calibration
Repeat Micro-Cases	Reinforce retention	Long-term learning

retained a 19.4 percentage point improvement above baseline after four weeks [12].

Retention varied by learner group. Radiology residents retained the most improvement, while medical students showed the largest decline from immediate post-test to follow-up. Written explanation quality also declined but remained above baseline. At follow-up, 47.9% of intervention comments identified a specific artifact and connected it to the model answer, compared with 22.5% at baseline.

Participants rated the intervention as useful. On a five-point scale, the mean usefulness rating was 4.4. Participants most often reported that paired examples helped them understand how model language could be influenced by visual artifacts. Several participants commented that they had previously thought of artifacts as a human interpretation issue but had not considered them as a model-output issue.

6. Discussion

The study shows that artifact literacy for medical vision-language model use can be taught in a short educational session. Participants who received artifact-specific training were better able to detect model answers that were influenced by artifacts, less likely to accept those answers without edits, and more likely to document specific uncertainty. The effect was larger than the improvement seen after a general AI caution module, which suggests that concrete image-based practice is more effective than broad warnings.

The findings also show that artifact literacy is not the same as general imaging experience. Radiology residents performed better at baseline than other groups, but they still improved after the intervention [13]. This matters because medical VLMs introduce a new interpretive layer [14]. A radiology trainee may recognize a skin fold or external marker but still need practice evaluating how a model has incorporated that feature into language. The learner must critique both the image and the model's explanation.

The improvement in written explanation quality is especially important. A participant who merely says "the AI may be wrong" has not provided a useful critique. A participant who says that the model's statement about a lesion border is uncertain because shadow and ruler placement distort the margin has shown artifact literacy. This specificity can support better teaching, feedback, and documentation. It also helps prevent vague skepticism from replacing careful evaluation.

The intervention reduced high-confidence inappropriate acceptance. This may be the most important safety-related educational outcome. Learners will not always detect every model error, but reducing confident acceptance of artifact-sensitive errors is valuable. Confidence can influence whether a learner challenges or repeats a model explanation. After training, participants were more cautious when artifacts were relevant, without becoming broadly dismissive of correct answers.

The modest increase in review time is acceptable for an educational setting. The goal of training is not to make learners faster immediately. It is to make them more accurate and more reflective. The fact that time increased mainly for artifact-related cases suggests that participants were applying additional scrutiny selectively. In future work, repeated practice may reduce the time cost as the checklist becomes habitual.

The partial decline at four weeks indicates that a single session is not enough for durable mastery. The intervention created measurable retention, but some gains faded. Artifact literacy should probably be reinforced through spaced practice. Short recurring cases during radiology conferences, emergency medicine teaching, der-

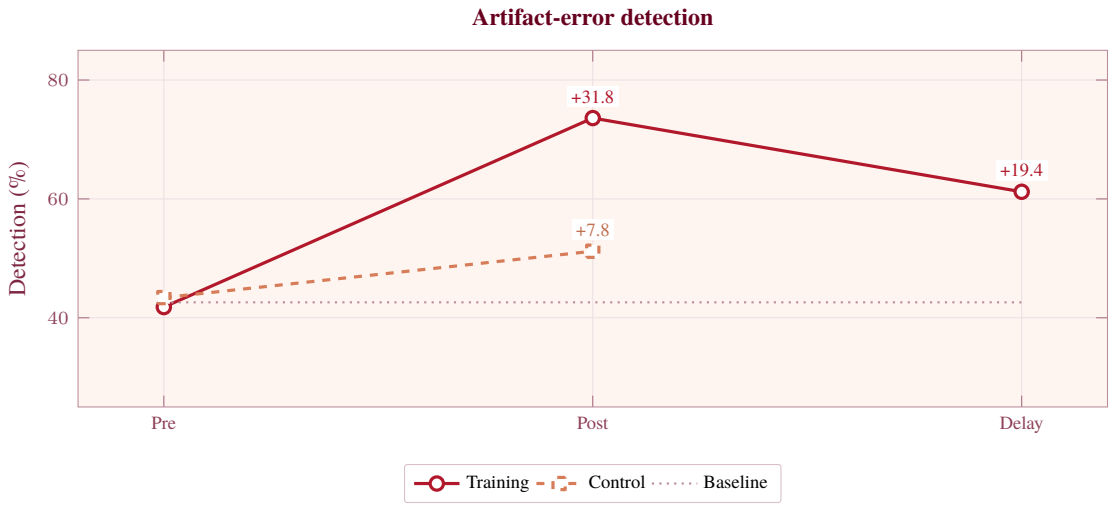

Figure 7: Primary outcome trajectory for artifact-related model error detection. The training group showed a large immediate gain and partial retention at the four-week delayed assessment.

matology image review, or clinical skills sessions may be more effective than a one-time workshop.

The modality results suggest that curricula should include both radiology and non-radiology images. Dermatology and ultrasound showed weaker baseline performance and substantial improvement. These domains may be particularly prone to lighting, angle, gain, and texture issues that influence model language. A curriculum limited to chest radiographs would not prepare learners for the broader range of image-based AI tools they may encounter.

The active control group improved slightly. This is expected because general caution can make participants more attentive. However, the control improvement was smaller and less specific. General AI literacy is useful, but artifact literacy requires visual practice. The study supports a layered curriculum: broad AI limitations first, followed by domain-specific error patterns and model-output critique.

The results also have implications for developers of educational platforms. If a platform presents model answers to learners, it should include cases where the model is wrong for artifact-related reasons. It should ask learners to identify the visual basis of the model's claim. It should provide feedback explaining whether the image feature is anatomical, external, technical, or uncertain. It should avoid presenting model output as a final answer without requiring evidence checking.

The study does not imply that learners should be trained to distrust medical VLMs entirely. Over-rejection increased only slightly, and the intervention emphasized calibrated review. Correct model answers can be educationally useful. The problem is uncritical acceptance. Artifact literacy teaches learners to ask when a model answer is supported, when it needs qualification, and when it should be rejected.

7. Educational Recommendations

Based on the study findings, artifact literacy should be treated as a discrete competency in medical AI education. It should not be hidden inside a generic lecture about limitations. A practical curriculum can be short, case-based, and repeated over time.

A useful training sequence should include the following elements:

- clean cases where the model answer is acceptable and well supported;
- cases where the model is wrong for ordinary diagnostic reasons;
- cases where an artifact visibly influences the model explanation;
- cases where an artifact is present but does not invalidate the model answer;
- cases requiring learners to revise the model answer rather than only accept or reject it.

The distinction between the third and fourth elements is important. If learners see only artifact-caused failures, they may begin rejecting any image with an artifact. Clinical imaging often contains artifacts, and many images remain interpretable. The educational objective is

Figure 8: Safety-sensitive response changes. Training sharply reduced inappropriate acceptance of artifact-influenced model answers while keeping over-rejection of acceptable answers below the predefined concern threshold.

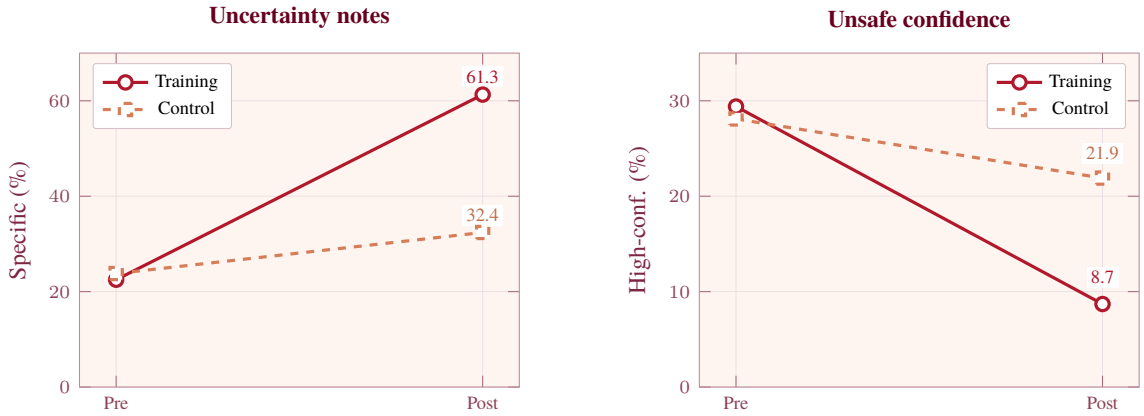

Figure 9: Documentation and calibration outcomes. The intervention increased specific uncertainty documentation and reduced the highest-risk behavior: confident acceptance of artifact-related model errors.

to judge the relationship between artifact, evidence, and model language.

A model-assisted image review checklist for learners should be brief enough to use during teaching:

- What is the model's main claim?
- What visible feature does the model use to support it?
- Is that feature anatomical, technical, external, or uncertain?
- Is the model more confident than the image permits?
- What qualification or correction should be added?

Educators should also teach learners to edit model responses. Editing is more realistic than binary acceptance. A model may identify the correct general finding but overstate certainty. It may describe a visible abnormality but use a misleading location. It may produce a useful first sentence followed by unsupported explanation. Learners need to practice preserving useful content while removing or qualifying unsupported content.

Assessment should include written justification. Multiple-choice acceptance decisions measure recognition, but written comments show whether learners can explain the problem. The scoring rubric used in this study can be adapted for educational programs. High-quality responses should identify the artifact and connect it to the model's claim.

Artifact literacy should be taught across specialties. Radiology programs are an obvious location, but emergency medicine, dermatology, surgery, primary care, and physician assistant education increasingly use medical im-

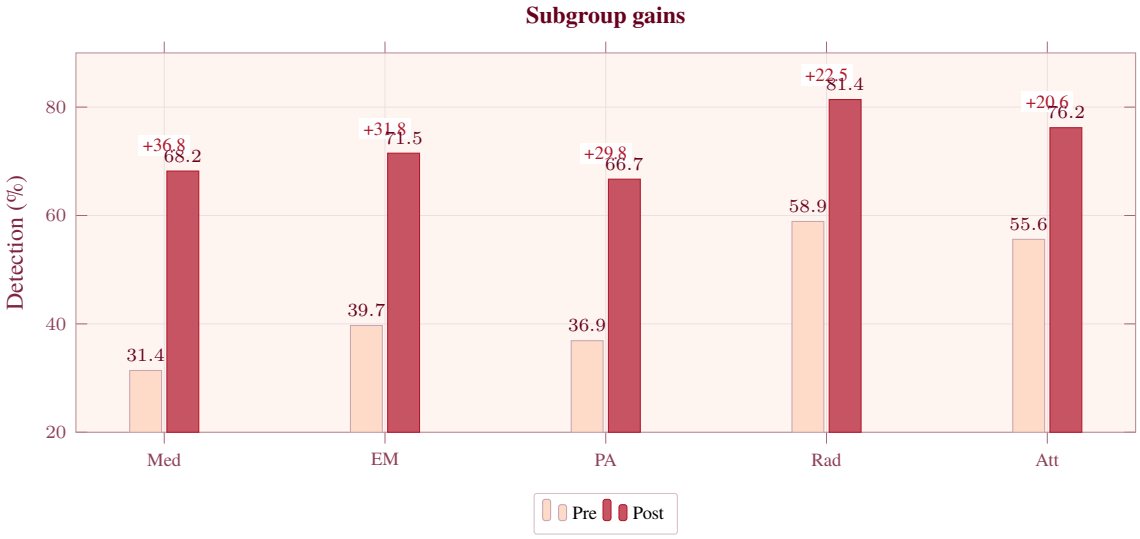

Figure 10: Training effects by learner group. The largest gains appeared among medical students and non-radiology trainees, while radiology residents and attending clinicians began from stronger baselines.

ages. Non-radiologists may be especially likely to encounter AI-generated image explanations in educational tools. They should be trained to recognize when the explanation depends on questionable visual cues.

The intervention should be repeated. A single session produced partial retention after four weeks, but skills declined. Spaced micro-cases could be added to existing teaching sessions. For example, a weekly conference could include one model-assisted case asking learners to identify whether the model's evidence is valid. Over time, this may normalize critical review.

8. Limitations

This study was conducted in simulated educational settings. Participants reviewed de-identified cases and did not make clinical decisions [15]. Performance in a real workflow may differ [16]. Clinical pressure, time constraints, incomplete information, and responsibility for patient care could change how learners use model outputs.

The intervention was brief. Its effect may depend on instructor quality, participant attention, and case selection. Although the module was standardized, live teaching can vary. Future studies should test asynchronous versions, longer curricula, and repeated sessions.

The participant sample came from teaching institutions. Results may not generalize to all hospitals, specialties, or levels of experience. The attending clinician subgroup was small. More research is needed among practicing clinicians who use imaging less frequently.

The cases were selected to represent artifact-related model behavior, not to estimate real-world frequency. The percentage of artifact-related model errors in this study should not be interpreted as the rate that would occur in clinical deployment. The dataset was designed for education and assessment.

The model responses were generated before the study and presented as fixed outputs. Participants did not interact dynamically with the model. In real use, learners may ask follow-up questions, and model answers may change. Interactive behavior may introduce new educational challenges.

The study measured four-week retention only. Longer follow-up is needed. It is possible that gains fade further without reinforcement. It is also possible that learners who continue seeing cases retain skills better.

The assessment rubric focused on artifact recognition and uncertainty documentation. It did not measure broader diagnostic accuracy in depth. A learner could improve artifact critique without improving overall image interpretation. Artifact literacy should complement, not replace, standard imaging education.

The curriculum used selected artifact categories. Other artifacts and image conditions may require additional training. Future modules should include more examples from pathology, ophthalmology, endoscopy, magnetic resonance imaging, and computed tomography [17].

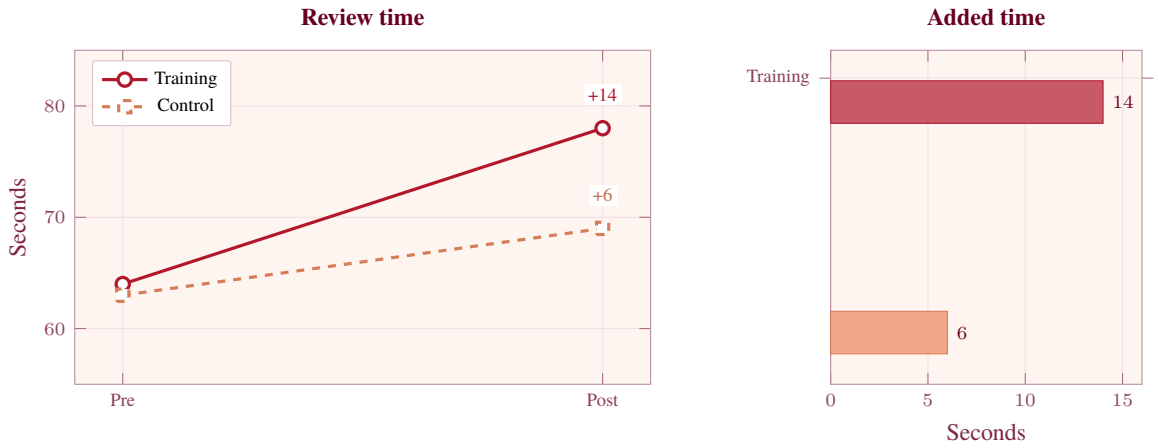

Figure 11: Efficiency outcome. Training increased average review time modestly, with the additional time concentrated in artifact-related cases rather than uniformly across all model-answer reviews.

9. Conclusion

This paper presented a controlled educational study of artifact literacy training for medical vision-language model use. The study showed that a short, case-based module improved participants' ability to detect artifact-related model errors, reduced inappropriate acceptance of flawed model explanations, improved written uncertainty documentation, and reduced high-confidence acceptance of artifact-sensitive outputs.

The intervention worked better than a general AI caution module, indicating that learners need concrete practice with image artifacts and model language. Broad warnings about AI limitations are not enough [18]. Learners benefit from seeing how a model's explanation can be shaped by external markers, image quality, overlays, lighting, shadow, gain, and other visual conditions.

The study also showed that artifact literacy is relevant across training levels and specialties. Radiology residents began with stronger skills, but all groups improved. Medical students and non-radiology trainees showed large gains, suggesting that this training may be especially useful early in medical education and in specialties where imaging is used but not the main focus of training.

The four-week follow-up showed that gains were partly retained but declined over time. Artifact literacy should therefore be reinforced through repeated practice. Educational programs that use medical vision-language models should include cases where model outputs are correct, incorrect, artifact-sensitive, and appropriately uncertain.

Medical vision-language models can be useful educational tools, but learners need preparation to use them critically. Artifact literacy offers a practical competency: inspect the image, identify possible artifacts, check the model's stated evidence, and qualify or reject the output when the evidence is weak. Teaching that competency can make model-assisted imaging education safer and more disciplined.

References

[1] C. S. Gray, A. Grudniewicz, A. Armas, J. W. Mold, J. Im, and P. Boeckxstaens, "Goal-oriented care: A catalyst for person-centred system integration.," *International journal of integrated care*, vol. 20, pp. 8–8, 11 2020.

[2] B. Sadanandan and V. Behzadan, "Vsf-med: A vulnerability scoring framework for medical vision-language models," *arXiv preprint arXiv:2507.00052*, 2025.

[3] I. N. Wong, O. Monteiro, D. T. Baptista-Hon, K. Wang, W. Lu, Z. Sun, S. Nie, and Y. Yin, "Leveraging foundation and large language models in medical artificial intelligence.," *Chinese medical journal*, vol. 137, pp. 2529–2539, 9 2024.

[4] T. R. Ahmad, A. W. Kong, M. L. Turner, J. Barnett, G. Kaur, K. S. O'Brien, N. D. Pasricha, and M. Indaram, "Socioeconomic correlates of keratoconus severity and progression.," *Cornea*, vol. 42, pp. 60–65, 2 2022.

[5] A. K. Sowan, A. R. Beraya, A. Carrola, C. Reed, S. V. Matthews, and T. Moodley, "Developing, implementing, and evaluating a multimedia patient de-

cision aid program to reform the informed consent process of a peripherally inserted central venous catheter procedure: Protocol for quality improvement.," *JMIR research protocols*, vol. 7, pp. e10709–, 12 2018.

[6] J. Chung, F. Aguila, and O. A. Harris, "Validity assessment of referral decisions at a va health care system polytrauma system of care.," *Cureus*, vol. 7, pp. e240–, 1 2015.

[7] D. R. Newman-Griffis, M. B. Hurwitz, G. P. McKernan, A. J. Houtrow, and B. E. Dicianno, "A roadmap to reduce information inequities in disability with digital health and natural language processing.," *PLOS digital health*, vol. 1, pp. e0000135–e0000135, 11 2022.

[8] C. J. Steer, P. R. Jackson, H. Hornbeak, C. K. McKay, P. Sriramarao, and M. P. Murtaugh, "Team science and the physician-scientist in the age of grand health challenges.," *Annals of the New York Academy of Sciences*, vol. 1404, pp. 3–16, 10 2017.

[9] F. A. Wilson, Y. Wang, and J. P. Stimpson, "Do immigrants underutilize optometry services," *Optometry and vision science : official publication of the American Academy of Optometry*, vol. 92, pp. 1113–1119, 11 2015.

[10] S. A. Martin and E. A. Frutiger, "Vision stations: Addressing corrective vision needs with low-cost technologies:," *Global advances in health and medicine*, vol. 4, pp. 46–51, 3 2015.

[11] N. N. Ludwig, D. T. Jashar, K. Sheperd, J. L. Pineda, D. Previ, J. Reesman, C. Holingue, and G. J. Gerner, "Considerations for the identification of autism spectrum disorder in children with vision or hearing impairment: A critical review of the literature and recommendations for practice.," *The Clinical neuropsychologist*, vol. 36, pp. 1049–1068, 12 2021.

[12] A. E. Barañano, J. Wu, K. Mazhar, S. P. Azen, and R. Varma, "Visual acuity outcomes after cataract extraction in adult latinos. the los angeles latino eye study.," *Ophthalmology*, vol. 115, pp. 815–821, 9 2007.

[13] R. L. Richesson, B. E. Bray, C. Dymek, R. A. Greenes, L. D. McIntosh, B. Middleton, G. Perry, J. Platt, and C. Shaffer, "Summary of second annual mcbk public meeting: Mobilizing computable biomedical knowledge—a movement to accelerate translation of knowledge into action," *Learning health systems*, vol. 4, pp. e10222–, 3 2020.

[14] M. Bunn, D. Khanna, E. Farmer, E. Esbrook, H. Ellis, A. Richard, and S. Weine, "Rethinking mental healthcare for refugees.," *SSM. Mental health*, vol. 3, pp. 100196–100196, 2 2023.

[15] D. S. W. Ting and A. Al-Aswad, "Augmented intelligence in ophthalmology: The six rights.," *Asia-Pacific journal of ophthalmology (Philadelphia, Pa.)*, vol. 10, pp. 231–233, 7 2021.

[16] A. Villagomez, F. M. Munoz, R. L. Peterson, A. M. Colbert, M. Gladstone, B. MacDonald, R. Wilson, L. Fairlie, G. Gerner, J. Patterson, N. S. Boghossian, V. J. Burton, M. Cortés, L. D. Katikaneni, J. C. G. Larson, A. S. Angulo, J. Joshi, M. Nesin, M. A. Padula, S. Kochhar, and A. K. Connery, "Neurodevelopmental delay: Case definition & guidelines for data collection, analysis, and presentation of immunization safety data.," *Vaccine*, vol. 37, pp. 7623–7641, 12 2019.

[17] W. Horner-Johnson, K. Dobbertin, J. C. Lee, and E. M. Andresen, "Disparities in chronic conditions and health status by type of disability," *Disability and health journal*, vol. 6, pp. 280–286, 6 2013.

[18] R. G. Baraniuk, D. L. Donoho, and M. Gavish, "The science of deep learning.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, pp. 30029–30032, 11 2020.