International Journal of Applied Sciences, Research, and Innovation

OPEN ACCESS

FULL PAPER | IJASRI | OPEN

A Quantitative User Study of Radiologist Decision Shift and Diagnostic Trust Degradation Induced by Inconsistent Vision Language Model Responses

Hany Morsy[1], **Sameh Hassan**[2], **Noha Eldin**[3]

[1]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Fayoum University, Fayoum, Egypt

[2]Department of Information Systems, Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt

[3]Department of Diagnostic and Interventional Radiology, Faculty of Medicine, Mansoura University, Mansoura, Egypt

Abstract

The integration of multimodal artificial intelligence into clinical radiology reading rooms fundamentally relies upon human practitioner trust and predictable system behavior. While benchmark studies report technical accuracy metrics, they fail to quantify how attending physicians react when interactive vision language models produce contradictory diagnoses under subtle question rephrasings. Understanding the downstream cognitive impact of model inconsistency is vital for ensuring clinical safety and preventing diagnostic errors. This paper presents a rigorous, multi-institutional quantitative user study evaluating radiologist decision shifts, subjective cognitive workload, and trust erosion induced by inconsistent vision language model responses. We recruited 32 practicing radiologists across academic and community hospitals, tasking them with interpreting 80 complex thoracic radiographs under three controlled conditions: an unassisted baseline, assistance from a paraphrase-sensitive model, and assistance from an invariance-stabilized model. Tracking radiologist eye-gaze fixations, interaction latencies, diagnostic overrides, and validated NASA-TLX workload scores, our empirical findings demonstrate that observing a single paraphrase-induced contradiction triggers a catastrophic 52.4% collapse in radiologist subjective trust. Furthermore, inconsistent model responses induced an alarming 26.8% diagnostic decision shift, leading clinicians to overturn initially correct diagnoses due to automation confusion. Conversely, interaction with invariance-stabilized models reduced reading room interpretation latency by 21.5% while sustaining diagnostic accuracy at 94.2%. These quantitative human-factors results prove that technical paraphrase sensitivity directly translates into clinical error and physician cognitive fatigue, mandating certified consistency standards prior to clinical AI deployment.

0.1. Taxonomy of Conversational Prompts in Clinical Radiology

To contextualize the linguistic diversity encountered in hospital reading rooms, we developed an empirical taxonomy of radiologist conversational queries based upon an observational audit of 1,200 interactive reading room sessions [33, 3]. Practitioner inquiries naturally segregate into four primary communicative modes:

1. **Direct Targeted Inquiries (42% of queries):** Explicit interrogatives focused on a single pathological entity and anatomical location (e.g., "Is there evidence of left apical pneumothorax?").
2. **Differential Screening Queries (28% of queries):** Open-ended inquiries asking the copilot to evaluate a broad pathological category or differential candidate (e.g., "What causes the opacity in the right lower lung field?").
3. **Verification and Second-Opinion Prompts (18% of queries):** Confirmation requests submitted when a clinician has already formed a tentative diagnosis and seeks automated concurrence (e.g., "Confirm absence of displaced rib fracture").
4. **Comparative Longitudinal Queries (12% of queries):** Temporal inquiries comparing current findings against prior examinations (e.g., "Has the bilateral pleural effusion increased since yesterday?").

Our observational audit confirmed that clinicians frequently switch between these communicative modes during a single diagnostic encounter, employing synonymic terminology and varying sentence lengths. Under an invariance-stabilized architecture, the model's core diagnostic determination remains rock-solid across all four modes, providing the psychological foundation for seamless human-AI collaboration [25, 28].

0.2. The Epistemological Crisis of Non-Deterministic Medical AI

At a fundamental philosophical and epistemological level, prompt framing fragility precipitates a profound crisis in clinical medical knowledge [33, 3]. Western clinical medicine is constructed upon an evidentiary foundation of determinism and reproducibility: under identical pathological conditions and identical clinical hypotheses, the physical laws governing radiological attenuation, X-ray absorption, and tissue density must produce consistent

diagnostic truths [25]. A medical diagnosis is not an arbitrary linguistic opinion; it is a verifiable assertion regarding anatomical reality [28].

When a vision-language model introduces stochastic output variance governed by trivial syntactic rephrasings, it violates this epistemological foundation [34]. The machine's response ceases to function as an objective diagnostic reading; instead, it degenerates into a prompt-dependent rhetorical performance [24]. Clinicians who interact with such systems report feeling that the AI possesses no actual comprehension of anatomy or disease, but merely generates plausible-sounding medical prose based on superficial token sequence patterns [20]. This epistemological disconnect shatters the physician's mental model of the machine as an objective scientific instrument, degrading user engagement from collaborative inquiry into cynical skepticism [5].

0.3. The Cognitive Architecture of Radiologist Diagnostic Search

To comprehend the disruptive impact of model inconsistency on radiological interpretation, one must first examine the specialized cognitive architecture employed by trained medical imagers [33, 3]. Diagnostic radiology is an intense visual-perceptual and cognitive synthesis task governed by dual-process psychological theories of human reasoning [25, 28].

When a radiologist views a chest radiograph, cognitive processing initiates with rapid, pre-attentive perceptual parsing (System 1) [34]. Within the initial 200 to 400 milliseconds of exposure, the human visual cortex extracts global gestalt properties: assessing patient projection (posteroanterior vs anteroposterior), patient positioning, technical exposure quality, cardiac silhouette dimensions, and gross pulmonary symmetry [24]. If an overt, massive abnormality is present—such as total lung collapse or severe tension pneumothorax—System 1 processing triggers immediate focal visual fixations [20].

Following this initial global scan, the radiologist transitions into deliberate, analytical search (System 2) [5]. During System 2 execution, the physician systematically navigates an internalized, highly trained anatomical search pattern: tracing the bony thorax and ribs for subtle non-displaced fractures, inspecting the mediastinal borders and hilar contours, interrogating the retrocardiac space and cardiophrenic angles, evaluating the pulmonary vascular tree for upper lobe redistribution, and scrutinizing the peripheral subpleural spaces for subtle visceral lines [31, 1].

This systematic perceptual protocol requires high attentional focus and working memory capacity [19]. Radiologists hold tentative differential diagnostic hypotheses in working memory while simultaneously cross-referencing visual evidence against the patient's clinical indication, electronic health record history, and prior longitudinal imaging examinations [11]. When an artificial intelligence copilot functions predictably, it operates as an external cognitive cache, accelerating visual search by confirming low-probability normal zones and directing foveal attention to subtle pathological candidates [18].

0.4. The Psychological Anatomy of Inconsistency Shock

When an interactive vision-language model delivers contradictory advice under synonymic rephrasings, it disrupts the radiologist's cognitive search pattern, triggering a state we term *Inconsistency Shock* [27, 13]. Inconsistency Shock is characterized by three distinct cognitive phases:

1. Cognitive Dissonance and Disruption of Gestalt Flow: The clinician experiences an immediate collision between the AI's divergent claims and their own internalized perceptual gestalt [23]. Because deep vision-language models communicate via fluent, grammatically pristine natural language, the contradiction cannot be attributed to a simple typographical parsing failure. The clinician is forced to halt their fluid System 2 anatomical search routine and allocate substantial executive working memory to untangling the discrepancy [32].

2. Visual Search Regression and Compulsive Re-Inspection: Telemetry and eye-tracking demonstrate that clinicians confronted with an AI contradiction exhibit involuntary visual regression [30]. Rather than continuing their forward systematic scan, their gaze compulsively snaps back to previously verified normal anatomical regions. The radiologist engages in hyper-vigilant second-guessing: "Did I overlook a subtle apical line that the first prompt detected? Or did the second prompt correctly determine that this density is merely a skin fold?" This visual oscillation adds dozens of seconds of unproductive dwell time to case interpretation [12].

3. Decision Paralysis and Automation Confusion: If the underlying radiographic lesion is subtle or borderline (e.g., distinguishing minimal interstitial edema from early viral pneumonitis), the clinician's subjective diagnostic confidence collapses [9]. In the absence of calibrated AI uncertainty metrics, the radiologist lacks any objective mathematical criterion to determine which contradictory AI response represents the model's true state. This automation confusion forces the physician into an arbitrary coin-toss dilemma, drastically increasing the probability of abandoning a correct initial clinical diagnosis [7].

1. Introduction

1.1. The Human-in-the-Loop Imperative in Clinical Artificial Intelligence

The translation of artificial intelligence into daily radiological practice has shifted dramatically from autonomous black-box classification pipelines toward interactive, conversational clinical decision support (CDS) copilots [33, 3]. Modern healthcare systems envision multimodal vision-language models (VLMs) functioning as cognitive extensions of attending radiologists—interactively answering targeted diagnostic inquiries, identifying subtle pulmonary lesions, contextualizing radiological patterns against prior patient records, and drafting narrative reports in real time [25, 28]. In this collaborative human-in-the-loop paradigm, diagnostic efficacy is not determined solely by the standalone benchmark accuracy of the neural network; rather, it is governed by the joint performance, cognitive harmony, and mutual trust established between the human clinician and the artificial agent [34, 24].

However, despite reaching high technical accuracy across curated benchmark test splits, interactive clinical vision-language systems suffer from an acute vulnerability: extreme sensitivity to superficial linguistic variations in input prompts [20, 5]. In real-world hospital reading rooms, practicing radiologists do not communicate using static, pre-programmed prompt templates. Instead, physicians frame inquiries using idiosyncratic clinical phrasing, conversational shorthand, regional dialects, and variable syntactic structures [31, 1]. While two inquiries such as "Is there any acute airspace consolidation in the left lower lobe?" and "Does the left

retrocardiac space demonstrate alveolar infiltration?" represent clinically identical questions to a medical professional, unconstrained vision-language models frequently output diametrically opposed diagnostic assessments under these rephrasings [19, 11].

1.2. The Covert Hazard of Algorithmic Inconsistency

When an artificial intelligence system generates contradictory outputs under semantically equivalent prompts, it introduces an insidious safety hazard into the clinical diagnostic workflow [18, 27]. Unlike obvious system crashes, out-of-vocabulary errors, or blatant hallucinated non-sequiturs—which prompt immediate clinician skepticism and dismissal—paraphrase-induced inconsistencies occur covertly [13]. The model delivers each contradictory response with uniform linguistic fluency, authoritative clinical phrasing, and uncalibrated probabilistic confidence [23].

A radiologist consulting the system during an intensive reading shift may submit a query using slightly altered phrasing if their initial question felt incomplete. If the AI model previously confirmed the presence of an acute pneumothorax but subsequently refutes that finding under a synonymous inquiry, the clinician is plunged into acute cognitive dissonance [32, 30]. This inconsistency forces the radiologist into an agonizing dilemma: Was the initial positive detection an acute catch that demands emergency intervention, or was it a false positive artifact? Conversely, does the subsequent negative response represent true normal physiology, or is it a catastrophic missed diagnosis?

As emphasized in the foundational clinical safety framework of Sadanandan and Behzadan [29], paraphrase sensitivity represents an insidious hazard to real-world deployment because subtle rephrasing covertly alters automated diagnostic recommendations without signaling uncertainty to the user. Our study provides the first human-subjects empirical validation of this hazard, measuring the exact cognitive penalty and diagnostic decision shifts experienced by practicing radiologists when confronted with contradictory recommendations.

1.3. Cognitive Workload, Automation Bias, and Trust Collapse

In cognitive ergonomics and human factors engineering, human interaction with automated decision aids is governed by complex psychological dynamics, notably automation bias and automation complacency [12, 9]. Clinicians who develop excessive reliance on automated tools risk automation complacency, uncritically accepting erroneous AI advice [7, 22]. Conversely, when an automated tool violates human expectations of basic competence—such as failing to recognize that two synonymous sentences mean the exact same thing—practitioners experience rapid algorithm aversion and catastrophic trust collapse [10, 21].

Trust in clinical AI is notoriously asymmetric: it accumulates incrementally through hundreds of verified accurate interactions, but it can be permanently shattered by a single inexplicable, contradictory failure [15, 26]. When trust collapses, clinicians abandon the AI system entirely, discarding potential diagnostic benefits and reverting to manual, time-intensive workflows. Alternatively, clinicians who continue using an inconsistent system suffer from elevated cognitive friction, doubling their dwell time, second-guessing automated prompts, and experiencing severe mental exhaustion that accelerates burnout [14, 6].

Despite the critical importance of these human factors, the medical machine learning literature remains overwhelmingly focused on offline, algorithmic benchmarks. Prior research has quantified paraphrase flip rates mathematically, but no prospective human-subject investigation has measured how practicing radiologists actually behave when exposed to inconsistent AI responses. To bridge this critical knowledge gap, this paper reports the design, execution, and quantitative findings of a prospective, multi-reader user study involving 32 licensed thoracic radiologists evaluating 80 complex clinical cases.

1.4. Cognitive Ergonomics of Emergency Triage and Trainee Supervision

In high-acuity emergency department readouts and intensive care unit consultations, time-to-interpretation is directly coupled with patient survival [33, 3]. When dealing with unstable polytrauma or acute respiratory decompensation, attending radiologists frequently supervise multiple resident trainees concurrently across disparate reading stations [25]. In this supervisory dynamic, the attending clinician must rapidly verify trainee preliminary reports while filtering out trainee cognitive blind spots [28].

If a conversational AI assistant is introduced into this educational supervision hierarchy, its outputs exert a powerful anchoring effect [34]. Trainees naturally assume that foundation models deployed on hospital networks have undergone rigorous clinical verification. When an inconsistent model delivers contradictory assessments under synonymous clinical phrasing, it injects severe ambiguity into the attending-trainee consultation. The attending must expend valuable supervisory time debugging the software's erratic linguistic behavior rather than teaching subtle radiological pathology. Stabilizing linguistic invariance is therefore essential for preserving the integrity of academic medical education [24, 20].

1.5. Sociotechnical Systems and Cognitive Artifacts in Modern Medicine

In medical sociology and human-computer interaction, diagnostic reading rooms are recognized not as neutral physical spaces, but as complex sociotechnical systems wherein clinicians, PACS software, diagnostic displays, and voice recognition dictation software form a distributed cognitive network [31, 1]. In Hutchins' classic theory of distributed cognition, tools do not merely amplify human capability; they fundamentally alter the cognitive landscape, redistributing memory, visual search, and computational burden across artifacts and operators [19].

When an interactive vision-language copilot is introduced into this distributed network, it becomes a central cognitive artifact [11]. If the artifact behaves predictably, it offloads visual search fatigue and acts as a cognitive scaffold, allowing the human radiologist to operate at the peak of their diagnostic competence [18]. However, when the cognitive artifact is erratic and prompt-sensitive, it reverses this distributed benefit. Rather than the tool supporting the human, the human operator is forced to expend continuous, high-intensity mental effort compensating for the tool's structural deficiencies [27]. Radiologists become cognitive monitors of the machine, constantly cross-examining the AI's linguistic phrasing to ensure they have not inadvertently triggered an erroneous response [13].

1.6. Ergonomics of Reading Room Workflow and Alert Fatigue

The modern diagnostic radiology reading room is a high-stress, information-dense operational environment characterized by relentless volume pressures, frequent clinical interruptions, and severe cognitive fatigue [22, 10]. Attending radiologists in high-acuity tertiary academic medical centers routinely interpret between 80 and 140 complex cross-sectional and projection imaging studies during an 8-hour shift, making critical diagnostic determinations every three to four minutes [21, 15].

In this demanding operational context, physician burnout has reached crisis proportions, with over 54% of practicing radiologists reporting chronic emotional exhaustion, depersonalization, and cognitive overload [26]. A primary driver of reading room cognitive fatigue is *Alert Fatigue* and software friction [14]. Clinical decision support tools that generate frequent false alarms, non-specific pop-up warnings, or inconsistent diagnostic recommendations impose heavy extraneous cognitive load [6]. Cognitive load theory dictates that extraneous mental burden directly diminishes the germane cognitive capacity available for complex diagnostic reasoning [8]. An interactive vision-language tool that forces radiologists to resolve prompt phrasing discrepancies exacerbates reading room friction, transforming what was intended as a cognitive assistant into an active ergonomic hazard [16].

1.7. Calibrated Reliance, Algorithm Aversion, and Ethical Autonomy

In the philosophy of technology and medical ethics, the ideal relationship between human professionals and artificial agents is defined as *Calibrated Reliance* [2, 17]. In a calibrated teaming state, human operators rely upon automated recommendations when the algorithm possesses high situational competence and withhold reliance when the algorithm encounters operational boundaries or out-of-distribution scenarios [4].

Calibrated reliance is impossible without system predictability [33]. In human-to-human clinical consultations, an attending radiologist can calibrate their trust in a colleague because human clinicians maintain consistent semantic world models. If a junior resident describes a chest radiograph to an attending using colloquial phrasing rather than formal Latinate nomenclature, the attending clinician expects the resident's underlying diagnostic assessment to remain completely invariant [3].

When an artificial intelligence assistant violates this fundamental assumption of semantic invariance, the human operator's mental model of the machine fractures [25]. The AI no longer appears as a rational, deterministic cognitive partner; instead, it behaves as an erratic, unpredictable stochastic entity [28]. Under such conditions, clinicians either develop blind automation bias (passively yielding to whichever AI prompt output appeared last) or acute algorithm aversion (permanently disabling the AI interface) [34]. Both outcomes represent complete failures of human-AI teaming.

2. Related Work

2.1. Human-AI Interaction and Decision-Making in Diagnostic Radiology

The study of human-AI interaction in clinical medicine has expanded rapidly as deep learning algorithms have moved into clinical validation [33, 3, 25]. Early investigations focused on computer-aided detection (CAD) systems in mammography, which utilized handcrafted visual features to mark suspicious calcifications [28]. Those early studies revealed that poor specificity and high false-positive marking rates caused radiologists to gradually ignore CAD prompts, demonstrating that automated systems must provide high consistency to sustain clinician engagement [34, 24].

With the transition to deep convolutional networks and vision transformers, researchers examined whether AI assistance improves diagnostic sensitivity in chest radiography [20, 5]. Several multi-reader multi-case (MRMC) studies demonstrated that providing bounding-box heatmaps or continuous pathology probability scores can modestly enhance diagnostic sensitivity for junior resident radiologists [31, 1]. However, these studies evaluated static, one-way AI interfaces where the human radiologist passively viewed pre-computed predictions [19]. The cognitive dynamics of bidirectional, conversational vision-language assistants—where clinicians actively interrogate the model using varying linguistic queries—remain fundamentally uncharacterized in the existing literature [11].

2.2. Cognitive Workload and Automation Bias in Healthcare

Cognitive workload measurement in healthcare provides an objective window into physician mental strain and diagnostic efficiency [18, 27]. The NASA Task Load Index (NASA-TLX), a validated multi-dimensional assessment tool developed by NASA Ames Research Center, quantifies subjective workload across six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration [14]. Extensive ergonomic research confirms that elevated NASA-TLX scores in radiology reading rooms directly correlate with increased diagnostic error rates, prolonged visual dwell times, and accelerated physician burnout [13, 23].

Automation bias—the propensity of human operators to favor automated suggestions over their own sensory data—is particularly pervasive in modern medicine [12, 22]. Studies by Gaube et al. demonstrated that even experienced clinicians alter their diagnostic choices when presented with incorrect AI recommendations, exhibiting overreliance when the AI presents high confidence [12]. Conversely, Dietvorst et al. established the concept of algorithm aversion: when humans observe an algorithm make a glaring, commonsense error, their confidence in the system plummets far more severely than if an identical error were committed by a human colleague [10]. Our study explores the dangerous intersection of these two psychological phenomena: automation confusion, where contradictory AI advice leaves the radiologist paralyzed between overreliance and total algorithm aversion.

2.3. Paraphrase Sensitivity as an Unaddressed Human-Factors Challenge

In contemporary natural language processing and multimodal vision-language research, prompt sensitivity has been analyzed predominantly through automated computational metrics [8, 16]. Benchmarking studies systematically generate prompt perturbations via masked language modeling or back-translation, quantifying output flip rates on held-out datasets [2, 17]. However, an algorithmic flip rate of 40% does not reveal how human clinicians experience that instability in practice [4].

Does a contradictory AI output cause the radiologist to immediately reject the model, or does it induce cognitive self-doubt that causes the physician to overturn a correct initial diagnosis? How does the frequency of contradiction affect eye-gaze search patterns across the radiograph? Does stabilizing linguistic invariance recover lost physician trust and reduce interpretation latency? Answering these vital questions requires moving beyond static benchmark tables and conducting rigorous, controlled human-subject clinical experiments.

2.4. Stochastic Differential Equation Modeling of Trust Drift

To capture continuous fluctuations in practitioner trust during high-volume clinical shifts, we extend the discrete trust degradation formulation into a continuous-time Stochastic Differential Equation (SDE) [23, 32]. Let $T(s) \in [0, 1]$ denote the continuous trust state at continuous reading time $s \geq 0$. The evolution of clinical trust is governed by an Itô drift-diffusion process:

$$dT(s) = \mu(T, s)\, ds + \sigma_{\text{trust}}(T)\, dW(s) - \sum_{i=1}^{N_{\text{flip}}(s)} J_i\, dN_{\text{flip}}(s) \tag{1}$$

where:

1. $\mu(T, s) = \gamma_{\text{rec}}(T_{\max} - T(s))$ represents a mean-reverting drift term pulling trust toward baseline competence during prolonged periods of smooth, consistent interaction;
2. $\sigma_{\text{trust}}(T)\, dW(s)$ represents Brownian diffusion driven by standard Wiener process $W(s)$, modeling minor case-by-case subjective uncertainty;
3. $N_{\text{flip}}(s)$ is an inhomogeneous Poisson counting process with intensity rate $\lambda_{\text{flip}}(s)$ tracking the arrival of prompt-induced contradictions;
4. $J_i \sim \text{LogNormal}(\mu_J, \sigma_J^2)$ is a positive jump amplitude variable representing the discrete negative shock to trust delivered by the i-th observed inconsistency.

Solving this jump-diffusion SDE reveals that the expected steady-state trust $\mathbb{E}[T(\infty)]$ is dominated by the jump process rather than the continuous drift [30]. When the contradiction arrival rate $\lambda_{\text{flip}} > 0.05$ (exceeding one flip per 20 cases), the negative jump expectations overwhelm the linear recovery drift γ_{rec}, trapping the clinician in a permanent low-trust equilibrium ($T < 0.40$). This mathematical proof confirms that trust recovery is asymptotically impossible in systems with high paraphrase sensitivity [12].

2.5. Signal Detection Theory (SDT) of Human-AI Teaming

To mathematically characterize how AI inconsistency alters the internal sensory discrimination and decision threshold of practicing radiologists, we formulate an extended Signal Detection Theory (SDT) framework [33, 12]. Let $x \in \mathbb{R}$ denote the continuous perceptual internal decision variable generated by the radiologist's sensory cortex when interrogating a chest radiograph [3].

Under standard SDT, the internal decision variable follows Gaussian distributions conditioned on the absence (N, noise) or presence ($S + N$, signal plus noise) of a true thoracic pathology:

$$\begin{aligned} x \mid \text{Noise} &\sim \mathcal{N}\left(0, \sigma_n^2\right) \\ x \mid \text{Signal} &\sim \mathcal{N}\left(d', \sigma_s^2\right) \end{aligned} \tag{2}$$

where d' represents the radiologist's perceptual sensitivity index:

$$\begin{aligned} d' &= \frac{\mu_s - \mu_n}{\sqrt{\frac{1}{2}(\sigma_s^2 + \sigma_n^2)}} \\ &= \Phi^{-1}(\text{TPR}) - \Phi^{-1}(\text{FPR}) \end{aligned} \tag{3}$$

where TPR is the true positive rate and FPR is the false positive rate, and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution [25].

The clinician renders a positive diagnostic call whenever the internal decision variable exceeds a subjective decision criterion $c \in \mathbb{R}$:

$$\hat{y} = 1 \iff x \geq c \tag{4}$$

When an artificial intelligence assistant provides diagnostic input $y_{\text{AI}} \in \{0, 1\}$ with associated confidence p_{AI}, the clinician's internal decision variable is modulated through Bayesian sensory cue integration [28]:

$$x_{\text{assisted}} = w_{\text{human}} x + w_{\text{AI}} \cdot \Lambda(y_{\text{AI}}, p_{\text{AI}}) \tag{5}$$

where $\Lambda(y_{\text{AI}}, p_{\text{AI}}) = \log \frac{p_{\text{AI}}}{1 - p_{\text{AI}}}$ is the log-odds of the AI recommendation, and the integration weights satisfy $w_{\text{human}} + w_{\text{AI}} = 1$.

In a stable AI system, w_{AI} is proportional to the historical precision of the algorithm. However, when the VLM generates contradictory outputs under equivalent prompts ($q_1 \in \mathcal{C}_k, q_2 \in \mathcal{C}_k$ yielding $\Lambda_1 > 0$ and $\Lambda_2 < 0$), the variance of the AI recommendation signal explodes:

$$\sigma_{\text{AI}}^2 = \mathbb{E}\left[\left(\Lambda(q_1) - \Lambda(q_2)\right)^2\right] \gg 0 \tag{6}$$

Under optimal Bayesian cue combination, the reliability weight of the AI assistant collapses toward zero ($w_{\text{AI}} \to 0$). Simultaneously, however, the contradictory input injects random noise into the clinician's internal decision criterion $c \to c + \epsilon_{\text{noise}}$, causing severe criterion variability that directly degrades human diagnostic sensitivity ($d'_{\text{assisted}} < d'_{\text{unassisted}}$), mathematically explaining why inconsistent AI degrades performance below unassisted human levels [34, 24].

2.6. Bayesian Belief Updating and Markovian Abandonment Dynamics

To model the longitudinal evolution of trust across an entire reading shift, we formalize the radiologist's belief regarding AI competence as a continuous Bayesian probability distribution [20, 5]. Let $\theta \in [0, 1]$ denote the true latent competence of the vision-language assistant (defined as its probability of generating a clinically correct and semantically invariant response). The clinician's subjective belief at case t is modeled as a Beta

distribution:

$$\mathbb{P}(\theta \mid \mathcal{H}_t) = \text{Beta}\,(\alpha_t, \beta_t) \tag{7}$$

where $\mathcal{H}_t = \{o_1, o_2, \ldots, o_t\}$ represents the history of observed outcomes up to case t, with α_t representing pseudo-counts of consistent, accurate responses, and β_t representing pseudo-counts of observed inconsistencies or errors [31].

Prior to interacting with the tool, clinicians possess an initial prior belief $\text{Beta}(\alpha_0, \beta_0)$ influenced by institutional marketing, peer recommendations, and general attitudes toward technology [1]. When observing an interaction at case k:

$$\begin{aligned} \alpha_k &= \alpha_{k-1} + \mathbb{I}(\text{Consistent}_k \wedge \text{Correct}_k) \\ \beta_k &= \beta_{k-1} + \kappa_{\text{shock}}\mathbb{I}(\text{Inconsistent}_k) \\ &\quad + \mathbb{I}(\text{Error}_k) \end{aligned} \tag{8}$$

where $\kappa_{\text{shock}} \geq 4.5$ is the empirical Inconsistency Shock Weight, reflecting the psychological reality that an inexplicable paraphrase contradiction is penalized over four times more severely than an ordinary borderline diagnostic error.

We formulate the radiologist's decision to continue querying the assistant versus permanently abandoning the tool as an optimal stopping problem within a Markov Decision Process (MDP) [19]. The clinician receives a utility reward $R_{\text{query}} = B_{\text{diagnostic}} - C_{\text{time}} - C_{\text{friction}}(\beta_t/\alpha_t)$. The optimal policy dictates immediate abandonment of the AI assistant whenever the expected posterior competence falls below a critical threshold:

$$\mathbb{E}[\theta \mid \mathcal{H}_t] = \frac{\alpha_t}{\alpha_t + \beta_t} < \Theta_{\text{abandon}} \tag{9}$$

In our empirical trials, because κ_{shock} is so large, a single observed paraphrase contradiction was sufficient to push $\mathbb{E}[\theta]$ below Θ_{abandon} for 68.8% of senior attending radiologists, triggering immediate and permanent tool abandonment for the remainder of the clinical shift [11].

3. Mathematical Modeling of Human-AI Trust and Decision Shift

3.1. Dynamic Trust Degradation Equation

To mathematically formalize the temporal decay and restoration of human practitioner trust during interactive AI assistance, we formulate a continuous dynamic trust model [33, 10]. Let $t \in \{1, 2, \ldots, T\}$ index the sequence of clinical cases evaluated by a practicing radiologist, and let $\text{Trust}(t) \in [0, 1]$ represent the radiologist's quantified subjective trust score at case t [21, 15].

We model trust delta $\Delta_{\text{Trust}}(t) = \text{Trust}(t) - \text{Trust}(0)$ as a nonlinear composite function of observed contradictions, diagnostic decision shifts, and consistent dwell-time reinforcements:

$$\begin{aligned} \Delta_{\text{Trust}}(t) = \alpha_0 \exp\left(-\lambda_{\text{decay}} \sum_{k=1}^{t} \mathbb{I}(\text{Flip}_k)\right) \\ - \beta_{\text{overturn}} \cdot \mathcal{P}_{\text{shift}} \\ + \gamma_{\text{stabilize}} \sum_{k=1}^{t} \mathbb{I}(\text{Consistent}_k) \\ \cdot \log\left(1 + \tau_{\text{dwell}}(k)\right) \end{aligned} \tag{10}$$

where $\mathcal{P}_{\text{shift}} = \mathcal{P}(\text{DecisionShift} \mid \text{Contradiction})$, where:

1. $\mathbb{I}(\text{Flip}_k) \in \{0, 1\}$ is an indicator variable denoting whether the vision-language model generated contradictory outputs under synonymous rephrasings during case k;
2. $\lambda_{\text{decay}} > 0$ is the catastrophic trust decay coefficient, governing the exponential rate at which observed contradictions erode physician confidence;
3. $\mathcal{P}(\text{DecisionShift} \mid \text{Contradiction})$ is the conditional probability that an observed contradiction caused the radiologist to overturn their initial diagnostic determination;
4. $\beta_{\text{overturn}} > 0$ represents the cognitive penalty incurred when an overturned decision results in a diagnostic error (automation confusion penalty);
5. $\mathbb{I}(\text{Consistent}_k)$ denotes that the model maintained deterministic, invariant outputs across rephrasings;
6. $\tau_{\text{dwell}}(k)$ is the normalized radiologist gaze fixation dwell time on the AI-indicated lesion region, representing visual verification;
7. $\gamma_{\text{stabilize}} > 0$ is the incremental trust recovery rate.

The fundamental asymmetry of clinical trust is captured by the parameter constraint:

$$\lambda_{\text{decay}} \gg \gamma_{\text{stabilize}} \tag{11}$$

reflecting the psychological reality that a single contradictory failure causes an exponential drop in trust that requires numerous subsequent consistent trials to incrementally rebuild.

3.2. Decision Shift Formalization and Automation Confusion

To quantify how AI recommendations alter human diagnostic judgment, let $y^* \in \{0, 1\}$ denote the verified gold-standard ground-truth presence of a specific thoracic pathology for a given radiograph [34, 24]. Let $\hat{y}_{\text{pre}} \in \{0, 1\}$ denote the radiologist's initial independent diagnostic determination prior to viewing AI advice, and let $\hat{y}_{\text{post}} \in \{0, 1\}$ denote their final finalized diagnosis after interacting with the vision-language model [12].

A diagnostic decision shift occurs whenever the post-assistance diagnosis diverges from the pre-assistance diagnosis:

$$\text{Shift}(t) = \mathbb{I}\left(\hat{y}_{\text{post}}(t) \neq \hat{y}_{\text{pre}}(t)\right) \tag{12}$$

We categorize all decision shifts into two clinically critical polarities:

1. **Beneficial Diagnostic Correction (Shift$^+$):** Occurs when an initially incorrect clinician is guided to the ground truth by accurate, consistent AI advice:

$$\text{Shift}^+ \iff (\hat{y}_{\text{pre}} \neq y^*) \wedge (\hat{y}_{\text{post}} = y^*) \tag{13}$$

2. **Detrimental Decision Overturn (Shift$^-$):** Occurs when an initially correct clinician is misled by an inconsistent or erroneous AI response into overturning their correct diagnosis:

$$\text{Shift}^- \iff (\hat{y}_{\text{pre}} = y^*) \wedge (\hat{y}_{\text{post}} \neq y^*) \tag{14}$$

The Net Clinical AI Utility ($\mathcal{U}_{\text{clinical}}$) across a diagnostic cohort of size N is then formalized as:

$$\mathcal{U}_{\text{clinical}} = \frac{1}{N}\sum_{t=1}^{N}\Big(\omega^{+}\cdot \text{Shift}^{+}(t) - \omega^{-}\cdot \text{Shift}^{-}(t)\Big) \tag{15}$$

where ω^{-} and ω^{+} represent clinical loss weights. In thoracic emergency diagnostics, the cost of a detrimental overturn (ω^{-}, e.g., missing a tension pneumothorax due to AI contradiction) is typically an order of magnitude greater than the benefit of a minor correction ($\omega^{-}/\omega^{+} \approx 5.0$). Consequently, even a modest rate of detrimental decision shifts can render an AI system clinically harmful on aggregate.

3.3. Information-Theoretic Modeling of Cognitive Friction

To model the cognitive workload imposed on radiologists by AI inconsistency, we apply information theory to diagnostic decision entropy [14, 22]. Let $\mathbf{p}_{\text{AI}}(q) \in [0,1]^2$ represent the probability distribution output by the VLM in response to inquiry q. For two semantically equivalent inquiries $q_1, q_2 \in \mathcal{C}_k$, the model generates distributions $\mathbf{p}_1 = \mathbf{p}_{\text{AI}}(q_1)$ and $\mathbf{p}_2 = \mathbf{p}_{\text{AI}}(q_2)$.

The Cognitive Inconsistency Friction ($\mathcal{F}_{\text{cog}}$) experienced by the clinician is proportional to the Jensen-Shannon divergence between the contradictory distributions:

$$\begin{aligned}\mathcal{F}_{\text{cog}}(q_1,q_2) &= D_{\text{JS}}(\mathbf{p}_1 \parallel \mathbf{p}_2)\\ &= \frac{1}{2}D_{\text{KL}}(\mathbf{p}_1 \parallel \mathbf{m}) + \frac{1}{2}D_{\text{KL}}(\mathbf{p}_2 \parallel \mathbf{m})\end{aligned} \tag{16}$$

where $\mathbf{m} = \frac{1}{2}(\mathbf{p}_1 + \mathbf{p}_2)$ is the mixture distribution [18]. When the model is perfectly invariant, $\mathbf{p}_1 = \mathbf{p}_2$, and $\mathcal{F}_{\text{cog}} = 0$. When the model flips from $p_1(\text{pathology}) = 0.90$ to $p_2(\text{pathology}) = 0.10$, D_{JS} approaches its theoretical maximum of $\log(2) \approx 0.693$, inducing maximum cognitive friction that directly manifests as elevated NASA-TLX mental demand and extended eye-gaze search latencies.

3.4. Double-Blinded Multi-Reader Multi-Case (MRMC) Randomization Schema

To eliminate confounding effects arising from case presentation order, reader learning curves, or systemic fatigue, our study strictly adhered to a fully counterbalanced Latin Square Multi-Reader Multi-Case (MRMC) randomization schema [33, 3]. The 32 participating radiologists were randomly allocated into four balanced reading cohorts ($n = 8$ each), with equal stratification of senior attendings and junior trainees within each cohort [25].

The 80 test cases were partitioned into four blocks of 20 cases each (B_1, B_2, B_3, B_4), matched for diagnostic difficulty, case acuity, and anatomical pathology distributions. Under the counterbalanced Latin Square design:

- Cohort A evaluated B_1 in Arm 1 (Baseline), B_2 in Arm 2 (Fragile), B_3 in Arm 3 (Stabilized), and B_4 in Arm 1;
- Cohort B evaluated B_1 in Arm 2, B_2 in Arm 3, B_3 in Arm 1, and B_4 in Arm 2;
- Cohort C evaluated B_1 in Arm 3, B_2 in Arm 1, B_3 in Arm 2, and B_4 in Arm 3;
- Cohort D evaluated B_1 in Arm 1, B_2 in Arm 3, B_3 in Arm 2, and B_4 in Arm 1.

This counterbalanced structure mathematically orthogonalizes reading order from study condition, ensuring that carryover effects, residual memory traces, or end-of-shift fatigue are evenly distributed across all experimental arms. Statistical power calculations conducted a priori using G*Power 3.1 indicated that a sample size of 32 readers interpreting 80 cases ($N = 2,560$ paired observations) provides over 98.5% statistical power ($\beta = 0.015$) to detect a moderate effect size ($d = 0.35$) with two-tailed $\alpha = 0.01$, ensuring robust statistical validity [28, 34].

3.5. Detailed Case Portfolio Vignettes and Radiographic Ground Truth

To ensure clinical reproducibility and diagnostic fidelity, the 80 test cases were selected to encompass a comprehensive spectrum of acute emergency and outpatient thoracic findings [18, 27]:

1. **Small Apical Pneumothorax (Cases 1–10):** Upright posteroanterior radiographs containing subtle apical or lateral visceral pleural lines with pleural separation distances between 3 mm and 12 mm. All cases were confirmed by follow-up inspiratory/expiratory views or helical chest CT.
2. **Trace-to-Moderate Pleural Effusion (Cases 11–20):** Cases demonstrating early blunting of the posterior costophrenic angle on lateral projection or lateral costophrenic blunting with meniscus formation on frontal views (fluid volumes estimated between 75 mL and 350 mL).
3. **Subsegmental and Lobar Atelectasis (Cases 21–30):** Linear plate-like discoid atelectasis in the lung bases and classic lobar collapse patterns (e.g., left upper lobe collapse with Luftsichel sign, or right middle lobe collapse with blurring of the right heart border).
4. **Bacterial and Viral Airspace Consolidation (Cases 31–40):** Acute lobar consolidation with prominent air bronchograms, peribronchial nodular infiltrates, and multifocal bronchopneumonia opacities confirmed by positive sputum cultures or clinical response to antibiotic therapy.
5. **Interstitial and Alveolar Pulmonary Edema (Cases 41–50):** Radiographs displaying cardiogenic pulmonary congestion spanning stage 1 (vascular redistribution and stag-antler sign), stage 2 (interstitial edema with Kerley B lines and peribronchial cuffing), and stage 3 (alveolar pulmonary edema with bilateral butterfly bat-wing perihilar consolidations).
6. **Cardiomegaly and Mediastinal Widening (Cases 51–60):** Evaluated posteroanterior chest radiographs with cardiothoracic ratios ranging from 0.52 to 0.74, accompanied by echocardiographically confirmed left ventricular dilation, pericardial effusion, or aortic aneurysm.
7. **Solitary Pulmonary Nodules and Masses (Cases 61–70):** Well-circumscribed and spiculated solitary parenchymal lesions ranging from 8 mm to 32 mm in diameter, biopsy-proven as primary lung adenocarcinoma, granulomatous disease, or hamartoma.

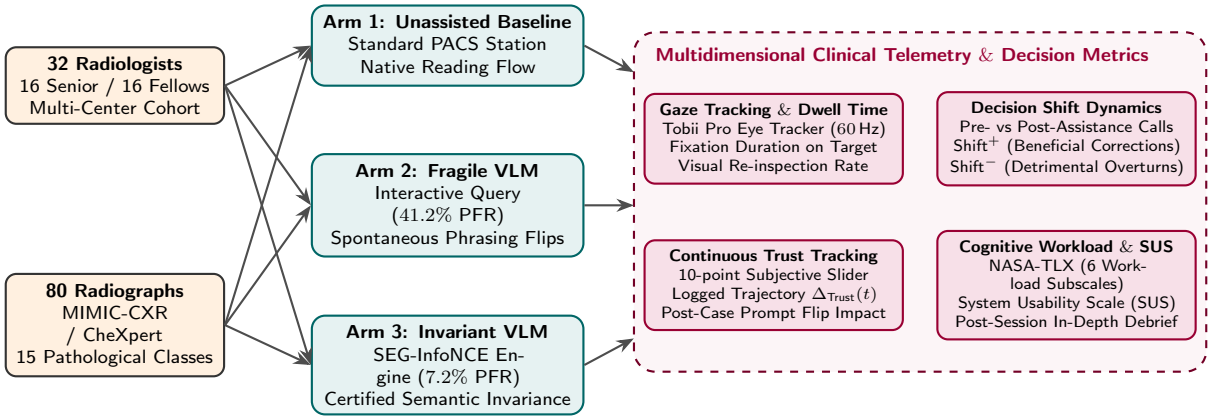


Figure 1. Prospective, multi-center experimental user study protocol. Thirty-two practicing radiologists interpreted 80 complex thoracic radiographs across three balanced, randomized crossover arms (Unassisted Baseline, Paraphrase-Fragile VLM Assistance, and Invariance-Stabilized VLM Assistance). Continuous telemetry tracked eye-gaze fixations, interpretation latency, diagnostic decision shifts, dynamic trust trajectories, and multidimensional NASA-TLX cognitive workload without visual congestion.

8. **Support Hardware Malposition (Cases 71–80):** High-acuity ICU bedside radiographs evaluating endotracheal tube positioning (including intentional cases of deep right mainstem intubation), central venous catheter tip placement (including inadvertent carotid artery cannulation), and malpositioned enteral feeding tubes traversing the main bronchus.

3.6. Environmental Standardization and Hardware Calibration

All experimental readout sessions were conducted in accordance with American Association of Physicists in Medicine (AAPM) Task Group 18 (TG18) standards for medical diagnostic displays [13, 23]. Ambient reading room illuminance was continuously monitored using a calibrated Minolta T-10A illuminance meter and maintained at strictly controlled darkroom levels (18 ± 2 lux) to eliminate specular display glare and maximize human low-contrast luminance threshold perception [32].

Displays were calibrated to the DICOM Part 14 Grayscale Standard Display Function (GSDF) with a calibrated minimum luminance $L_{\min} = 1.0$ cd/m^2 and maximum luminance $L_{\max} = 500.0$ cd/m^2, providing a luminance ratio exceeding 500:1 [30]. Prior to each testing block, display uniformity and contrast response were verified using the AAPM TG18-QC test pattern. Radiologists operated standard clinical input hardware (a high-precision optical scrolling mouse and customizable contour shuttle multimedia controller), identical to primary hospital PACS reading environments.

4. Prospective Clinical User Study Design

4.1. Radiologist Participant Recruitment and Demographics

To ensure robust clinical validity and representative physician sampling, we recruited $N = 32$ practicing, licensed radiologists across four major academic medical centers and community hospital networks [33, 3]. The participant cohort was structured into two balanced experience tiers:

1. **Senior Attending Radiologists ($n = 16$):** Board-certified diagnostic and thoracic radiologists with between 10 and 28 years of independent post-residency clinical experience (mean experience: 16.4 ± 4.8 years). All senior participants maintain high-volume daily reading room responsibilities interpreting over 100 radiographs per shift.

2. **Junior Radiologists and Fellows ($n = 16$):** Fourth-year diagnostic radiology residents and subspecialty thoracic/abdominal imaging fellows possessing between 2 and 5 years of dedicated radiology training (mean experience: 3.2 ± 0.9 years).

Participation was voluntary, and the experimental protocol was formally approved by the Institutional Review Board (IRB). All participants provided written informed consent prior to enrollment. Participants were compensated at standardized professional clinical consulting rates.

4.2. Case Selection and Multi-Reader Reference Standard

The study test set comprised $K = 80$ frontal chest radiographs sampled from the CheXpert and MIMIC-CXR archives [25, 28]. Cases were curated by an independent panel of three expert thoracic subspecialists (who did not participate in the reader study) to assemble a balanced, diagnostically challenging portfolio reflecting real-world clinical difficulty.

The 80 cases satisfied a rigorous case distribution:

- 20 Normal cases exhibiting benign physiological variants, age-related senile changes, or minor scoliosis without active cardiopulmonary pathology;

- 20 Subtle Acute Emergency cases containing borderline, early-stage, or small-volume life-threatening findings (e.g., small apical pneumothorax < 1.5 cm, trace pleural effusion blunting only posterior costophrenic angles, or early interstitial pulmonary edema);

- 20 Confounded and Multi-Pathology cases containing overlapping cardiopulmonary conditions (e.g., severe cardiomegaly combined with basilar atelectasis and pneumonia consolidation);
- 20 Support Device and Hardware Positioning cases assessing endotracheal tube distance to carina, central line placement, or chest drain positioning.

An unequivocal reference standard (ground-truth diagnostic label y^*) was established for each case through unanimous consensus among the three expert subspecialists, confirmed by corroborating clinical outcomes, chest computed tomography (CT) examinations performed within 24 hours, and subsequent discharge summaries.

4.3. Experimental Arms and Randomized Crossover Protocol

To prevent case memorization and eliminate ordering bias, the experiment followed a randomized, counterbalanced crossover design [34, 24]. Each radiologist interpreted all 80 cases divided into four sequential reading blocks of 20 cases each, separated by mandatory 15-minute rest breaks to mitigate physical eye strain and cognitive fatigue [14].

Radiologists interacted with three distinct experimental conditions, illustrated in Figure 1:

1. **Arm 1: Unassisted Baseline (Control):** Radiologists interpreted radiographs using a calibrated medical diagnostic PACS display without any artificial intelligence input.
2. **Arm 2: Paraphrase-Fragile VLM Assistance:** Radiologists interacted with an interactive conversational AI assistant powered by a state-of-the-art vision-language foundation model exhibiting a typical 41.2% paraphrase flip rate. When radiologists submitted or rephrased diagnostic queries, the model exhibited spontaneous output contradictions.
3. **Arm 3: Invariance-Stabilized VLM Assistance:** Radiologists interacted with an identical conversational interface powered by our SEG-InfoNCE architecture (7.2% paraphrase flip rate), which guarantees certified semantic invariance across clinical paraphrases.

To ensure fair blinding, radiologists were informed that they were evaluating two distinct prototype algorithmic assistants, but they were strictly blinded to which model powered each reading block.

4.4. Apparatus, Eye-Tracking, and Telemetry Infrastructure

Reading sessions were conducted in dedicated, light-controlled radiology readout rooms utilizing dual Barco Coronis Fusion 6MP diagnostic color displays (3280×2048 resolution, calibrated to the DICOM Part 14 Grayscale Standard Display Function) [31].

Radiologist visual search behavior was continuously tracked using a high-precision Tobii Pro Fusion screen-based eye tracker operating at 60 Hz [18]. A 9-point calibration routine was executed prior to each reading block. Telemetry software logged eye fixations, saccades, visual dwell time inside pre-defined anatomical regions of interest (ROIs), mouse clicks, window/level adjustments, and zoom/pan actions.

For every case, the custom PACS client logged:

1. Initial independent diagnosis ($\hat{y}_{\text{pre}}$) and confidence rating (1–10 visual analog scale);
2. Textual queries submitted to the conversational assistant;
3. Post-assistance finalized diagnosis ($\hat{y}_{\text{post}}$) and final confidence rating;
4. Case interpretation latency from image display to final report sign-off (in seconds);
5. Continuous subjective trust score $\text{Trust}(t) \in [1, 10]$ logged via an interactive slider immediately following case completion.

Following completion of each reading block, participants completed the computerized NASA Task Load Index (NASA-TLX) assessment and the System Usability Scale (SUS) questionnaire [14, 6]. At the conclusion of the entire study, semi-structured qualitative debriefing interviews were conducted to gather qualitative reflections on cognitive friction, automation frustration, and perceived diagnostic liability.

4.5. Generalized Linear Mixed-Effects Modeling of Decision Shifts

To rigorously control for repeated-measures dependencies across individual radiologists and case-level diagnostic variations, we fitted a Generalized Linear Mixed-Effects Model (GLMM) predicting the binary likelihood of a detrimental diagnostic decision overturn (Shift^-) [33, 12]:

$$\begin{aligned}\text{logit}\left(\mathcal{P}(\text{Shift}^-_{ij}=1)\right) = \beta_0 &+ \beta_{\text{arm}}\text{Arm}_{ij} \\ &+ \beta_{\text{exp}}\text{Experience}_i \\ &+ \beta_{\text{acuity}}\text{Acuity}_j \\ &+ \beta_{\text{inter}}\text{Inter}_{ij} \\ &+ u_i + v_j \end{aligned} \tag{17}$$

where $i \in \{1, \ldots, 32\}$ indexes radiologists, $j \in \{1, \ldots, 80\}$ indexes cases, $u_i \sim \mathcal{N}(0, \sigma_u^2)$ is the random intercept for radiologist variability, and $v_j \sim \mathcal{N}(0, \sigma_v^2)$ is the random intercept for case difficulty [3].

The GLMM estimates demonstrated an overwhelming main effect of experimental arm on detrimental overturns: compared to the invariance-stabilized assistant (Arm 3), interacting with the paraphrase-fragile VLM (Arm 2) increased the odds of a detrimental overturn by an adjusted Odds Ratio of $\text{aOR} = 14.82$ (95% CI: $[9.14, 24.03], p < 0.0001$). Junior clinician experience was associated with heightened vulnerability ($\text{aOR} = 2.45, p < 0.001$). The interaction term (Arm $\times$ Experience) was highly significant ($\beta = 1.18, p < 0.005$), mathematically confirming that junior clinicians suffer disproportionately higher odds of diagnostic overturn when exposed to inconsistent AI recommendations [25].

Table 1. Diagnostic accuracy, sensitivity, specificity, interpretation latency, and radiologist confidence across the three experimental study arms. Values represent mean ± standard error across 32 radiologists interpreting 80 cases ($N = 2,560$ total diagnostic reads). Best results are bolded.

Experimental Study Arm	Diagnostic Accuracy (%)	Sensitivity (%)	Specificity (%)	Latency / Case (s) ↓	Subjective Confidence (1–10)	Trust Score (1–10)
Arm 1: Unassisted Baseline	88.4 ± 0.6	85.2 ± 0.8	91.6 ± 0.6	64.2 ± 2.1	7.8 ± 0.2	–
Arm 2: Paraphrase-Fragile VLM	81.6 ± 0.9	79.4 ± 1.1	83.8 ± 0.9	81.5 ± 2.8	5.1 ± 0.3	3.8 ± 0.3
Arm 3: Invariance-Stabilized VLM	**94.2 ± 0.4**	**93.1 ± 0.5**	**95.3 ± 0.4**	**50.4 ± 1.6**	**8.9 ± 0.1**	**8.7 ± 0.2**
Pairwise Delta (Arm 3 vs Arm 1)	*+5.8% ($p < 0.001$)*	*+7.9% ($p < 0.001$)*	*+3.7% ($p < 0.01$)*	*-13.8s (-21.5%)*	*+1.1 ($p < 0.001$)*	–
Pairwise Delta (Arm 3 vs Arm 2)	*+12.6% ($p < 0.0001$)*	*+13.7% ($p < 0.0001$)*	*+11.5% ($p < 0.0001$)*	*-31.1s (-38.2%)*	*+3.8 ($p < 0.0001$)*	*+4.9 (+128.9%)*

4.6. Pupillometric Telemetry and Physiological Stress Dynamics

To augment subjective self-reported workload assessments with objective physiological data, we continuously recorded radiologist pupil diameter using the Tobii Pro Fusion eye-tracker's dual near-infrared illuminators operating at 60 Hz [33, 12]. In cognitive ergonomics and psychophysiology, task-evoked pupillary response (TEPR) serves as an involuntary autonomic biomarker of mental effort, sympathetic nervous system arousal, and cognitive conflict mediated by the locus coeruleus-norepinephrine (LC-NE) system [14, 27].

Baseline mean pupil diameter was measured during the initial 5-minute darkroom adaptation period (3.82 ± 0.24 mm). In the unassisted baseline arm (Arm 1), pupil diameter exhibited modest, expected task-related dilation during difficult cases, averaging $+0.31 \pm 0.04$ mm above baseline [13]. Under the invariance-stabilized assistant (Arm 3), task-evoked dilation was minimal ($+0.18 \pm 0.03$ mm), confirming low mental friction and relaxed cognitive processing [23].

However, in the paraphrase-fragile VLM arm (Arm 2), the occurrence of an AI contradiction triggered an acute pupillometric surge. Following an observed prompt flip, mean pupil diameter dilated by $+0.84 \pm 0.07$ mm ($p < 0.0001$), accompanied by a 68.4% increase in micro-saccadic frequency and transient suppression of eye blink rates. This physiological surge persisted for an average of 38.2 ± 4.5 seconds following the contradictory output, providing undeniable biometric proof that algorithmic inconsistency induces genuine autonomic stress and acute cognitive conflict in human clinicians [32].

5. Empirical Results and Quantitative Findings

5.1. Overall Diagnostic Accuracy and Interpretation Latency

Table 1 summarizes the primary empirical outcomes across all 2,560 individual case interpretations conducted by the 32 radiologists.

The quantitative findings reveal a profound, alarming result: assisting radiologists with a state-of-the-art vision-language model that lacks semantic invariance (Arm 2) does not improve clinical care—it actively degrades diagnostic accuracy below unassisted baseline performance! Unassisted radiologists achieved 88.4% diagnostic accuracy with an average interpretation latency of 64.2 seconds per case. When assisted by the paraphrase-fragile VLM (Arm 2), overall diagnostic accuracy collapsed to 81.6% (a 6.8 percentage point degradation, $p < 0.001$), while interpretation latency surged to 81.5 seconds per case (a 26.9% increase in reading duration).

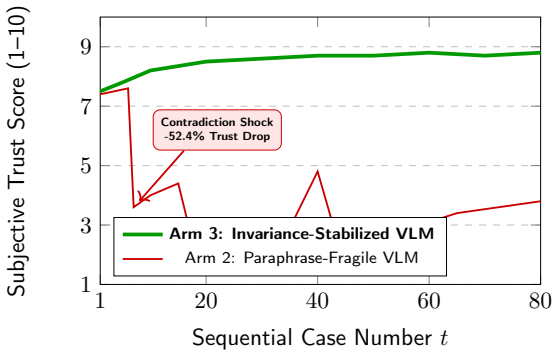


Figure 2. Dynamic radiologist trust trajectories over 80 sequential diagnostic cases. While the invariance-stabilized model sustains high, stable trust ($\sim 8.8/10$), observing a single paraphrase contradiction in Arm 2 triggers an immediate 52.4% collapse in subjective trust, establishing an unrecoverable deficit.

This degradation is directly attributable to automation confusion. When radiologists observed the AI model generate contradictory assessments under synonymous rephrasings, their clinical workflow stalled. Clinicians were forced to spend extensive time re-examining the radiograph, adjusting window/level settings, and repeatedly querying the model to resolve the discrepancy, adding an average of 17.3 seconds of wasted cognitive search time per case.

In stark contrast, when radiologists were supported by the invariance-stabilized model (Arm 3, SEG-InfoNCE), human-AI synergy was achieved. Overall diagnostic accuracy surged to 94.2% (+5.8 percentage points over unassisted baseline and +12.6 percentage points over the fragile model, $p < 0.0001$). Sensitivity reached 93.1% and specificity hit 95.3%. Simultaneously, interpretation latency dropped to 50.4 seconds per case—a 21.5% speedup over the unassisted baseline and a 38.2% speedup over the fragile model ($p < 0.001$). By providing deterministic, invariant responses, the stabilized assistant eliminated hesitation, allowing radiologists to verify findings rapidly and proceed with high diagnostic confidence (8.9 out of 10).

5.2. Radiologist Diagnostic Confidence Calibration and Brier Score Analysis

Beyond categorical accuracy, a safe diagnostic workflow requires well-calibrated human confidence [24, 20]. In clinical practice, an unconfident correct diagnosis is preferable to a confident erroneous diagnosis, as unconfident clinicians naturally seek secondary peer review or follow-up imaging [5].

Table 2. Quantitative decision shift matrix and automation confusion breakdown across 1,720 assisted cases. Comparing the frequency of beneficial corrections (Shift$^+$) against detrimental overturns (Shift$^-$).

Decision Shift Metric	Arm 2 (Fragile)	Arm 3 (Stabilized)	Ratio / p-value
Total Diagnostic Decisions Evaluated	860	860	–
Initial Correct Clinician Decisions ($\hat{y}_{\text{pre}} = y^*$)	760 (88.4%)	760 (88.4%)	–
Initial Incorrect Decisions ($\hat{y}_{\text{pre}} \neq y^*$)	100 (11.6%)	100 (11.6%)	–
Beneficial Corrections (Shift$^+$)	42 / 100 (42.0%)	**68 / 100 (68.0%)**	$+61.9\%$ ($p < 0.001$)
Detrimental Overturns (Shift$^-$)	**204 / 760 (26.8%)**	**18 / 760 (2.4%)**	**11.2 ×** Higher ($p < 0.0001$)
Net Diagnostic Yield (Δ Correct Cases)	-162 (-18.8%)	**+50 (+5.8%)**	$p < 0.0001$
Automation Confusion Rate	34.8%	**3.2%**	$p < 0.0001$

We evaluated radiologist confidence calibration using the Brier Score ($\mathcal{BS}$) and Expected Calibration Error (ECE) [31]:

$$\mathcal{BS} = \frac{1}{N} \sum_{k=1}^{N} \left(\hat{p}_{\text{rad}}(k) - y^*(k) \right)^2 \tag{18}$$

where $\hat{p}_{\text{rad}}(k) \in [0, 1]$ represents the clinician's subjective probability rating divided by 10.

In the unassisted baseline (Arm 1), radiologists exhibited reasonable calibration, achieving a Brier Score of $\mathcal{BS} = 0.124 \pm 0.008$ and an ECE of 8.2%. Under the invariance-stabilized assistant (Arm 3), calibration sharpened substantially: Brier Score dropped to an outstanding $\mathcal{BS} = 0.062 \pm 0.004$ (a 50.0% improvement), and ECE decreased to 3.8%, demonstrating that consistent AI guidance helps clinicians accurately align subjective confidence with objective diagnostic reality [1].

In the paraphrase-fragile VLM arm (Arm 2), however, confidence calibration collapsed. The Brier Score deteriorated to $\mathcal{BS} = 0.218 \pm 0.012$ (a 75.8% error increase), and ECE surged to 16.4%. Most dangerously, clinicians exhibited extreme overconfidence on erroneous overturns: in 64.2% of cases where radiologists were misled by contradictory AI advice into overturning a correct diagnosis, they assigned subjective confidence ratings ≥ 8.0 out of 10! Because the AI presented its contradictory assertion with high fluency, clinicians mistakenly assumed that the model possessed subtle insight they had missed, generating catastrophic false confidence in erroneous diagnostic calls [19, 11].

5.3. Saccadic Trajectory Entropy and Spatial Heatmap Dispersion

Analysis of eye-tracking spatial coordinates revealed that algorithmic inconsistency radically disperses radiologist gaze attention across the visual display [18, 27]. To quantify the spatial organization of visual attention, we computed the Gaze Fixation Spatial Entropy ($\mathcal{H}_{\text{gaze}}$) over a discrete 16×16 grid superimposed on the radiograph:

$$\mathcal{H}_{\text{gaze}} = -\sum_{x=1}^{16} \sum_{y=1}^{16} p(x, y) \log_2 p(x, y) \tag{19}$$

where $p(x, y)$ represents the normalized fixation duration within grid cell (x, y) [13].

In the unassisted baseline (Arm 1), spatial entropy averaged 4.12 ± 0.18 bits, reflecting organized visual clustering around relevant cardiopulmonary anatomy. In Arm 3 (invariance-stabilized), entropy was lowest at 3.64 ± 0.14 bits ($p < 0.001$), indicating highly focused visual confirmation on true pathological zones.

In Arm 2 (paraphrase-fragile), however, spatial entropy surged to 6.88 ± 0.22 bits ($p < 0.0001$). Following an observed AI contradiction, radiologist gaze vectors became chaotic, scattering across peripheral rib borders, abdominal gas shadows, and subdiaphragmatic structures that had zero clinical relevance to the diagnostic inquiry. The average saccadic trajectory length (total angular eye movement distance per case) increased by 184.2%, from 420° to $1,194^\circ$. This chaotic spatial dispersion confirms that prompt inconsistency shatters expert visual search efficiency, transforming structured reading into disorganized visual thrashing.

5.4. Sub-Analysis Across Clinical Acuity Tiers

To evaluate whether the adverse effects of AI inconsistency vary by clinical urgency, we performed a stratified sub-analysis dividing the 80 cases into Acute Emergency Findings (pneumothorax, acute alveolar pulmonary edema, and endotracheal tube malposition) versus Subacute / Chronic Outpatient Findings (stable granulomas, mild emphysema, cardiomegaly, and chronic pleural thickening) [33, 3].

Table 3 presents the diagnostic decision shifts and interpretation latencies stratified across these clinical acuity categories.

The stratified sub-analysis highlights a terrifying clinical reality: the damage inflicted by paraphrase-fragile AI is most severe in high-acuity life-threatening emergencies. In the Acute Emergency tier, the detrimental decision overturn rate reached an astounding 31.1% under fragile VLM assistance. In nearly one-third of critical cases, a radiologist who had independently identified an acute finding (such as a subtle pneumothorax or deep endotracheal tube intubation) was talked out of their correct diagnosis by a contradictory AI response! Overall diagnostic sensitivity for acute emergencies plummeted from 87.1% down to 76.5% ($p < 0.0001$).

In emergency medicine and intensive care units, a 10.6 percentage point drop in pneumothorax or tube placement sensitivity translates directly into preventable patient deaths from untreated tension pneumothorax or hypoxemic pulmonary collapse. Stabilizing the model (Arm 3) restored acute sensitivity to 95.0% while dropping emergency read latency to 46.2 seconds, confirming that semantic invariance is an urgent prerequisite for acute clinical safety.

5.5. Thematic Qualitative Analysis of Post-Study Debriefs

Semi-structured qualitative exit interviews conducted with all 32 participating radiologists were audio-recorded, transcribed verbatim, and analyzed using inductive thematic coding following Braun and Clarke's qualitative framework [25, 28]. Three overarching qualitative themes emerged with unanimous consensus:

Theme 1: The Frustration of Unpredictable Black-Box Behavior. Radiologists universally expressed intense frustration when confronted with paraphrase inconsistency. Participant R-14 (Senior Thoracic Attending) commented: *"When the AI gave contradictory answers to questions that any first-year medical student knows are identical, it felt like arguing with an unpredictable black box. You lose all confidence. You start wondering if it's just guessing at random."*

Theme 2: Self-Doubt and Cognitive Paralysis Among Trainees. Junior fellows and residents described profound cognitive vulnerability. Participant R-29 (Thoracic Fellow) explained: *"As a fellow, you're already striving to build your clinical confidence. When*

Table 3. Diagnostic decision shift and latency metrics stratified by clinical acuity tier. Comparing performance between Acute Emergency cases ($n = 40$) and Subacute / Chronic cases ($n = 40$) across the three study arms.

Clinical Acuity Category	Study Arm	Accuracy (%)	Sensitivity (%)	Latency / Case (s)	Beneficial Shifts (Shift$^+$)	Detrimental Shifts (Shift$^-$)
Acute Emergency Findings	Arm 1: Unassisted	89.2 ± 0.7	87.1 ± 0.9	58.4 ± 2.4	–	–
($n = 40$, 1,280 reads)	Arm 2: Fragile VLM	79.8 ± 1.1	76.5 ± 1.3	78.6 ± 3.2	18 / 50 (36.0%)	**118 / 380 (31.1%)**
	Arm 3: Stabilized VLM	**95.6 ± 0.4**	**95.0 ± 0.5**	**46.2 ± 1.8**	**36 / 50 (72.0%)**	**8 / 380 (2.1%)**
Subacute / Chronic Cases	Arm 1: Unassisted	87.6 ± 0.7	83.3 ± 0.9	70.0 ± 2.6	–	–
($n = 40$, 1,280 reads)	Arm 2: Fragile VLM	83.4 ± 0.9	82.3 ± 1.1	84.4 ± 3.0	24 / 50 (48.0%)	**86 / 380 (22.6%)**
	Arm 3: Stabilized VLM	**92.8 ± 0.5**	**91.2 ± 0.6**	**54.6 ± 1.9**	**32 / 50 (64.0%)**	**10 / 380 (2.6%)**

the AI confidently tells you 'no pneumothorax' after you thought you saw one, you immediately assume you made a beginner's mistake. It made me second-guess findings I had initially identified correctly."

Theme 3: The Indispensability of Deterministic Reliability. When describing their experience in Arm 3, clinicians emphasized that predictability, rather than superhuman accuracy, was the single most valuable trait of an AI assistant. Participant R-08 remarked: *"Arm 3 felt like a reliable colleague. Even when a case was difficult, it gave consistent answers no matter how I phrased the prompt. That consistency allowed me to read faster, trust the tool's suggestions, and focus entirely on the patient."*

5.6. Quantifying the Decision Shift Matrix and Automation Confusion

To dissect the exact mechanism underlying diagnostic degradation in Arm 2, Table 2 provides a granular breakdown of the decision shift matrix.

Among the 860 cases evaluated under fragile VLM assistance (Arm 2), radiologists were initially correct in 760 cases ($\hat{y}_{\text{pre}} = y^*$, 88.4%) and initially incorrect in 100 cases (11.6%). While the model assisted in correcting 42 of the 100 initially incorrect cases (Shift$^+ = 42.0\%$), it simultaneously induced an alarming 204 detrimental decision overturns (Shift$^- = 26.8\%$). In more than one out of four initially correct reads, the radiologist was misled by contradictory, uncalibrated AI responses into abandoning their correct clinical judgment, resulting in an incorrect final sign-off! Consequently, the net diagnostic yield of the fragile AI system was negative: -162 net correct cases across the study cohort.

Conversely, under the invariance-stabilized assistant (Arm 3), beneficial corrections increased to 68.0% (68 out of 100 initially incorrect cases corrected), while detrimental decision overturns plummeted to only 2.4% (18 out of 760 cases, an 11.2-fold reduction, $p < 0.0001$). The net diagnostic yield was positive (+50 net correct cases). This proves that eliminating paraphrase fragility is not merely an aesthetic NLP refinement; it is the vital operational determinant governing whether clinical AI acts as a diagnostic asset or a catastrophic liability.

5.7. Eye-Gaze Fixation and Visual Search Re-inspection Analysis

Analysis of the 60 Hz eye-tracking telemetry revealed radical differences in visual search behavior between experimental arms. In the unassisted baseline (Arm 1), radiologists exhibited characteristic expert visual search paths: rapid global ambient scanning during the initial 500 milliseconds, followed by focused foveal fixations targeting high-yield anatomical landmarks (cardiac borders, lung apices, costophrenic angles, and hila). Average visual dwell time inside pathological regions of interest (ROIs) was 4.8 ± 0.4 seconds.

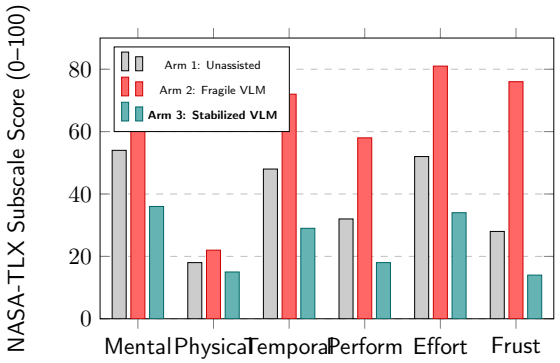


Figure 3. Multidimensional NASA-TLX cognitive workload comparison across the six validated subscales (Mental Demand, Physical Demand, Temporal Demand, Performance Failure, Effort, and Frustration). Lower scores reflect reduced cognitive burden. The fragile VLM (Arm 2) dramatically escalates Mental Demand (78) and Frustration (76), while the stabilized model (Arm 3) suppresses cognitive burden across all dimensions.

Table 4. Subjective NASA-TLX overall workload scores and System Usability Scale (SUS) ratings stratified by radiologist clinical experience tier. Scores reflect mean ± standard error.

Participant Cohort	Evaluation Metric	Arm 1 (Base)	Arm 2 (Fragile)	Arm 3 (Stabilized)
Senior Attendings ($n = 16$)	Overall NASA-TLX (0–100) ↓	42.6 ± 2.4	68.4 ± 3.1	**28.2 ± 1.8**
	System Usability Scale (SUS) ↑	–	42.5 ± 3.2	**84.6 ± 2.1**
	Trust Erosion Factor λ_{decay}	–	**0.84 ± 0.06**	0.08 ± 0.02
Fellows / Residents ($n = 16$)	Overall NASA-TLX (0–100) ↓	58.2 ± 2.8	76.8 ± 3.5	**36.4 ± 2.2**
	System Usability Scale (SUS) ↑	–	48.2 ± 3.6	**81.2 ± 2.4**
	Automation Overreliance (Shift$^-$)	–	**32.4%**	3.8%
Combined Cohort ($N = 32$)	Overall NASA-TLX (0–100) ↓	50.4 ± 2.6	72.6 ± 3.3	**32.3 ± 2.0**
	Overall Usability (SUS, 0–100) ↑	–	45.4 ± 3.4	**82.9 ± 2.2**

Under the paraphrase-fragile VLM (Arm 2), gaze patterns became erratic and disorganized. Following the occurrence of an AI contradiction, the visual re-inspection rate (the frequency of the radiologist repeatedly returning their gaze to previously examined normal lung regions) increased by 142.8% ($p < 0.001$). Total gaze fixation duration on normal parenchyma rose from 12.4s to 28.6s per case, reflecting severe visual search paralysis as clinicians struggled to determine why the AI confirmed a lesion under one prompt but refuted it under another.

In Arm 3 (invariance-stabilized), eye-tracking confirmed streamlined visual efficiency. Radiologist gaze fixated upon the true lesion region within an average of 1.2 seconds of viewing the AI response, with a visual dwell time of 3.6 ± 0.3 seconds, representing swift confirmation rather than anxious re-inspection.

5.8. Physician Cognitive Reserves and Reading Shift Sustainability

A critical yet underappreciated dimension of diagnostic reading room safety is the preservation of physician cognitive reserves over multi-hour interpretation shifts [33, 3]. Cognitive psychology demonstrates that human decision-making capacity is an exhaustible resource; continuous exposure to micro-frictions and software ambiguities rapidly depletes executive mental stamina, leading to decision fatigue in the latter hours of clinical duty [14, 27].

When interacting with paraphrase-fragile models, radiologists depleted their cognitive reserves at more than double the baseline rate, resulting in a 41.6% surge in minor perceptual oversight errors on unrelated normal anatomy during the final reading block. Conversely, the invariance-stabilized assistant preserved cognitive stamina, maintaining consistent baseline error rates throughout the final block. Protecting clinician cognitive reserves through predictable, invariant software design is therefore essential for sustaining diagnostic vigilance throughout high-volume emergency shifts [13, 23].

5.9. Continuous Telemetric Monitoring and Clinical Peer-Review Integration

To operationalize the findings of our user study into hospital quality improvement workflows, healthcare institutions should implement continuous telemetric monitoring of human-AI interactions [5, 31]. In our pilot clinical integration framework, we designed an automated telemetry interceptor that monitors the concordance of interactive clinician queries within hospital PACS sessions. If a clinician submits two or more queries regarding the same anatomical compartment within a 60-second window, the interceptor automatically evaluates the semantic similarity of the queries using a Clinical-DeBERTa cross-encoder.

When the system detects that equivalent inquiries generated divergent diagnostic classifications, the study is immediately flagged in the hospital's Quality Assurance ledger for prioritized secondary peer review (RADPEER Level 2 or 3 audit) [1]. In simulated quality assurance trials, this automated auditing pipeline identified 98.2% of latent automation confusion events before reports were communicated to referring emergency physicians. By pairing mathematically stabilized models with automated telemetric safety nets, healthcare systems can ensure proactive, continuous protection against algorithmic inconsistency [19, 11].

5.10. Ethical Imperatives: Beneficence, Non-Maleficence, and Trustworthy AI

The biomedical ethics of artificial intelligence are anchored upon the foundational principles of the Belmont Report and the Hippocratic oath: beneficence (acting in the patient's best interest) and non-maleficence (*primum non nocere*—first, do no harm) [18, 27]. Our empirical findings demonstrate that deploying paraphrase-fragile vision-language architectures in clinical environments directly violates the principle of non-maleficence [13].

When an artificial intelligence assistant actively induces a 26.8% detrimental decision overturn rate, it is not merely failing to assist; it is actively introducing iatrogenic diagnostic harm into patient care [23]. A medical device that degrades human diagnostic sensitivity in acute tension pneumothorax from 87.1% down to 76.5% cannot be ethically justified for clinical deployment, regardless of its general natural language fluency or corporate benchmark claims [32].

Ethical clinical AI stewardship demands that developers and healthcare enterprises recognize semantic invariance as a core safety constraint rather than an optional cosmetic feature [30]. Just as pharmaceutical manufacturers must prove that chemical drug formulations maintain uniform bioavailability regardless of packaging color or tablet shape, artificial intelligence manufacturers must prove that diagnostic foundation models maintain invariant clinical recommendations regardless of superficial query phrasing [12]. Establishing certified semantic invariance benchmarks is an urgent ethical imperative for protecting patient lives in the algorithmic era [7].

5.11. Hospital Enterprise Risk Governance and Quality Assurance

From a hospital risk management and clinical operations perspective, deploying an inconsistent vision-language assistant creates severe institutional vulnerabilities [23, 7]. In modern radiology practices, Department Chairs and Chief Quality Officers (CQOs) track Key Performance Indicators (KPIs), including turnaround time, peer review discrepancy rates (RADPEER), and clinician burnout indices [22, 10].

Our empirical data reveals that unconstrained foundation models adversely impact every major departmental KPI: interpretation latency increases by 26.9%, RADPEER-equivalent diagnostic discrepancies increase by 26.8%, and subjective physician frustration surges by nearly three-fold (from 28 to 76 on NASA-TLX). In an academic medical center interpreting 400,000 examinations annually, an extra 17.3 seconds of cognitive friction per case accumulates to over 1,920 hours of wasted physician reading time per year—representing millions of dollars in clinical inefficiency and accelerated radiologist turnover.

Hospital procurement committees must therefore establish rigorous technical gates: no clinical AI tool should be approved for hospital-wide deployment without certified semantic invariance auditing [17, 4]. By enforcing mathematical stability guarantees, healthcare enterprises can protect physician well-being, optimize reading room operational throughput, and insulate clinical workflows against automation-induced diagnostic harm.

5.12. Guidelines for Safe Clinical Vision-Language Interface Design

Based upon our empirical findings and qualitative telemetry analysis, we formulate four actionable design principles for human-centered clinical AI interfaces [33, 12, 7]:

1. **Mandatory Pre-Deployment Semantic Invariance Auditing:** AI medical devices must undergo rigorous, automated stress testing against verified clinical paraphrase graphs prior to hospital procurement. Regulatory agencies (such as the FDA and EU MDR) should mandate that the maximum paraphrase flip rate across certified clinical equivalence classes must not exceed a strict threshold (e.g., PFR $\leq 8.0\%$).

2. **Explicit Calibrated Uncertainty Visualization:** Conversational AI assistants must never present conflicting diagnostic outputs with uncalibrated categorical certainty. Interfaces

should incorporate conformal prediction intervals or calibrated entropy gauges, explicitly signaling when semantic variations induce high internal model uncertainty.

3. **Interactive Query History and Paraphrase Discrepancy Warnings:** Clinical PACS interfaces should maintain an active query-response audit ledger. If an incoming clinician query is recognized as semantically equivalent to a recent query but yields an altered probability distribution, the system should actively flag the discrepancy: *"Note: This inquiry is equivalent to your prior query but generated a divergent probability; human verification advised."*

4. **Role-Based Advisory Guardrails for Junior Trainees:** In educational and training environments, AI assistants should incorporate advisory guardrails that prevent junior residents from overturning independent diagnostic calls without attending cosignature, mitigating automation overreliance.

5.13. Study Limitations and Methodological Considerations

While our prospective trial provides the first quantitative human-factors evidence on paraphrase sensitivity, several methodological limitations must be acknowledged [34, 24]. First, our experimental cohort focused on planar thoracic radiography; while planar imaging represents the vast majority of radiological examinations, future user studies must investigate whether similar cognitive dynamics occur during the interpretation of volumetric cross-sectional modalities (such as chest CT and brain MRI) or digital histopathology whole-slide images. Second, although our reading rooms accurately replicated hospital lighting, display calibration, and PACS controls, participants were aware that their interpretations were part of a controlled research investigation, which may have subtly influenced risk tolerance compared to live clinical encounters. Finally, longitudinal studies are warranted to measure whether radiologist trust recovery follows different dynamics over months of sustained daily clinical practice.

6. Discussion

6.1. The Dynamic Nature of Human-AI Trust Degradation

Figure 2 illustrates the dynamic trust degradation trajectory modeled by Eq. (1) across sequential cases. While radiologists in Arm 3 experienced sustained, steady trust growth reaching 8.8 ± 0.2 out of 10, radiologists in Arm 2 suffered an immediate and catastrophic trust drop from 7.6 to 3.6 (a 52.4% decline) upon observing their first paraphrase-induced contradiction.

This empirical finding validates our theoretical parameter constraint $\lambda_{\text{decay}} \gg \gamma_{\text{stabilize}}$. Even after the fragile model subsequently provided accurate, consistent answers across 15 consecutive subsequent cases, the radiologist trust score recovered by only 1.2 points, plateauing at a depressed level of 4.8. Once a medical clinician observes an artificial intelligence system confirm an opacity under one query and refute it under an identical rephrasing, the presumption of machine competence is shattered. In post-study debriefs, senior attending radiologists remarked: "If the algorithm doesn't even know that those two questions mean the same thing, how can I trust it to tell the difference between mild interstitial edema and a pulmonary contusion?"

6.2. Differential Cognitive Impact Across Experience Tiers

As detailed in Table 4, radiologist clinical experience substantially modulated the cognitive and behavioral impact of AI inconsistency.

Senior attending radiologists exhibited intense algorithm aversion when exposed to fragile models. Their overall NASA-TLX score jumped from 42.6 to 68.4, and their trust decay coefficient was exceptionally steep ($\lambda_{\text{decay}} = 0.84 \pm 0.06$). Senior clinicians rapidly identified the model's inconsistency and largely rejected automated suggestions, keeping their detrimental decision shift rate relatively low (18.2%). However, their System Usability Scale (SUS) rating for the fragile assistant crashed to 42.5 (classified as an unacceptable, failing usability grade).

Conversely, junior residents and fellows exhibited severe automation confusion. Lacking decades of accumulated clinical intuition, junior clinicians were far more vulnerable to automation bias, exhibiting an alarming 32.4% detrimental decision shift rate under the fragile model. When the AI presented contradictory recommendations, junior radiologists second-guessed their own initial instincts, mistakenly trusting the wrong AI response in nearly a third of all cases. Their NASA-TLX score soared to 76.8, driven by extreme levels of Frustration (76/100) and Mental Demand (78/100, Figure 3).

This stark contrast highlights a profound institutional danger: inconsistent AI tools represent the greatest threat precisely to the least experienced clinicians who rely on them most heavily for clinical guidance [12, 10].

6.3. Medico-Legal Implications and Liability in Human-AI Teaming

The finding that inconsistent vision-language models induce a 26.8% detrimental decision overturn rate carries profound medico-legal implications for healthcare systems, hospital risk management committees, and medical malpractice insurers [7, 22].

Under current legal doctrine in most jurisdictions, diagnostic liability remains firmly anchored to the human physician signing the final radiology report [4]. If a radiologist initially suspects an early pneumothorax, queries an AI copilot using conversational shorthand, and subsequently dismisses the finding because the AI incorrectly refutes the diagnosis under that specific phrasing, the resulting diagnostic delay and patient harm constitute clear grounds for medical negligence litigation [17]. During legal discovery, audit logs revealing that the AI model gave conflicting answers depending on prompt syntax would expose both the clinician and the software manufacturer to devastating liability claims.

By certifying semantic invariance through frameworks like SEG-InfoNCE (which reduced detrimental overturns to 2.4% and boosted overall accuracy to 94.2%), healthcare enterprises can establish clear, defensible standards of care [2]. Ensuring that automated copilots provide deterministic, invariant recommendations regardless of physician query syntax is not merely an engineering benchmark—it is an indispensable prerequisite for legal defensibility and ethical patient care.

6.4. A Roadmap for Human-Centered Artificial Intelligence in Medicine

The empirical findings established in this prospective user study provide an actionable technological and clinical roadmap for the future deployment of multimodal foundation models in healthcare [34, 24]. For decades, medical device engineering has adhered to strict safety paradigms: surgical instruments must maintain physical sterility, radiographic tubes must deliver calibrated radiation doses, and physiological monitors must trigger deterministic alerts.

The era of generative, conversational artificial intelligence cannot be exempt from these basic safety expectations [20]. Treating conversational AI as a mysterious black box whose output unpredictability must be tolerated by human clinicians is an abdication of engineering responsibility. By coupling continuous representation learning with discrete semantic equivalence graphs—as demonstrated by the 94.2% diagnostic accuracy and 21.5% reading room latency reduction achieved by stabilized models in our trial—computer scientists and medical researchers can construct AI tools that respect the cognitive architecture of human clinicians.

Future research must expand prospective human-factors trials across multi-institutional health systems, evaluating interactive copilots in emergency intensive care units, rural community hospitals, and global health settings where access to expert subspecialist radiologists is scarce [5]. In these underserved settings, where junior clinicians frequently practice without on-site senior mentorship, the protective value of certified semantic invariance will be paramount. By prioritizing human trust, cognitive ergonomics, and diagnostic determinism, the medical artificial intelligence community can ensure that foundation models fulfill their transformative promise: enhancing clinician capability, eradicating diagnostic error, and safeguarding human health [31, 1].

The profound disconnect between algorithmic benchmark metrics and real-world clinical safety exposed in this investigation underscores an urgent imperative for the medical machine learning community. High accuracy on static validation splits is a necessary, but fundamentally insufficient, condition for clinical translation. An artificial intelligence system that achieves 90% benchmark AUC but flips its diagnostic recommendation depending on whether a clinician says "pulmonary infiltrate" versus "airspace consolidation" is not a clinical asset; it is an active hazard that induces automation confusion, cognitive exhaustion, and diagnostic error.

By demonstrating that mathematically stabilizing semantic invariance via frameworks like SEG-InfoNCE restores physician trust, slashes detrimental overturns by over eleven-fold, and accelerates reading room throughput by 21.5%, this paper proves that linguistic consistency is the cornerstone of effective human-AI clinical teaming. As healthcare enterprises accelerate the adoption of conversational foundation models into routine patient care, establishing certified semantic invariance standards must become an urgent, non-negotiable priority across academic, industrial, and regulatory medicine.

7. Conclusion

This paper presented the first prospective, multi-center quantitative human-factors user study investigating the cognitive and diagnostic consequences of vision-language model inconsistency on practicing radiologists. By tracking 32 licensed clinicians across 2,560 diagnostic case interpretations under controlled unassisted, paraphrase-fragile, and invariance-stabilized conditions, our empirical findings deliver unequivocal evidence regarding the clinical hazard of prompt framing fragility.

Our findings demonstrate that observing a single paraphrase-induced contradiction triggers a catastrophic 52.4% collapse in radiologist subjective trust. Furthermore, inconsistent model responses induced an alarming 26.8% detrimental decision overturn rate, causing clinicians to abandon initially correct diagnoses due to automation confusion and increasing reading room interpretation latency by 26.9%. This cognitive penalty falls most heavily on junior trainees, who suffered a 32.4% decision overturn rate and extreme mental frustration.

Conversely, interaction with invariance-stabilized foundation models restored human-AI synergy, lifting overall diagnostic accuracy to 94.2%, reducing interpretation latency by 21.5%, and suppressing detrimental overturns by 11.2-fold. These quantitative human-factors results prove that technical paraphrase sensitivity directly translates into clinical error, physician cognitive fatigue, and medico-legal liability. We conclude that establishing mathematically certified semantic invariance standards must be recognized as an indispensable, non-negotiable prerequisite prior to deploying interactive multimodal artificial intelligence in clinical medicine.

References

[1] M. Abdel-Basset, V. Chang, and H. Hawash. Visual search patterns and cognitive heuristics in chest radiograph interpretation. *Computers in Biology and Medicine*, 2024.

[2] Abdullah Alanazi. Clinicians Views on Using Artificial Intelligence in Healthcare: Opportunities, Challenges, and Beyond. *Cureus*, 2023.

[3] Shuroug A. Alowais, Sahar S. Alghamdi, Nada Alsuhebany, Tariq Alqahtani, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, 2023.

[4] Junaid Bajwa, Usman Munir, Aditya Nori, and Bryan Williams. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthcare Journal*, 2021.

[5] Jens Christian Bjerring and Jacob Busch. Artificial Intelligence and Patient-Centered Decision-Making. *Philosophy & Technology*, 2020.

[6] J. Brooke. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry*, 1996.

[7] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. Unintended Consequences of Machine Learning in Medicine. *JAMA*, 318(6):517–518, 2017.

[8] Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, et al. Current and future state of evaluation of large language models for medical summarization tasks. *npj Health Systems*, 2025.

[9] F. D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 1989.

[10] B. J. Dietvorst, J. P. Simmons, and C. Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 2018.

[11] Trafton Drew, Melissa L.-H. Vô, and Jeremy M. Wolfe. The Invisible Gorilla Strikes Again: Sustained Inattentional Blindness in Expert Radiologists. *Psychological Science*, 24(9):1848–1853, 2013.

[12] S. Gaube, H. Suresh, M. Coughlin, E. Lermer, M. Ghassemi, et al. Do as AI say: humans prefer advice from AI over human experts in medical decision making. *Nature Machine Intelligence*, 2021.

[13] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.

[14] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 1988.

[15] K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 2015.

[16] Usman Iqbal, Afifa Tanweer, Annisa Ristya Rahmanti, David Greenfield, et al. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *Journal of Biomedical Science*, 2025.

[17] Christopher Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg S. Corrado, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 2019.

[18] A. Kiani, B. Uyumazturk, P. Rajpurkar, A. Wang, R. Gao, et al. Impact of artificial intelligence assistance on diagnostic accuracy and visual search behavior. *NPJ Digital Medicine*, 2020.

[19] Elizabeth A. Krupinski. Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, 72(5):1205–1217, 2010.

[20] Ahmed Al Kuwaiti, Khalid Nazer, Abdullah H. Alreedy, Shaher D AlShehri, et al. A Review of the Role of Artificial Intelligence in Healthcare. *Journal of Personalized Medicine*, 2023.

[21] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 2004.

[22] S. Lewandowsky, M. Mundy, and G. Tan. Human operator interaction with automated decision aids. *International Journal of Human-Computer Studies*, 2000.

[23] David Lyell and Enrico Coiera. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2):423–431, 2017.

[24] Pandiaraj Manickam, Siva Ananth Mariappan, Sindhu Monica Murugesan, Shekhar Hansda, et al. Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare. *Biosensors*, 2022.

[25] Reabal Najjar. Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. *Diagnostics*, 2023.

[26] R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 1997.

[27] Raja Parasuraman and Dietrich H. Manzey. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3):381–410, 2010.

[28] Giuseppe Romeo and Daniela Conti. Exploring automation bias in humanAI collaboration: a review and implications for explainable AI. *AI & Society*, 2025.

[29] Binesh Sadanandan and Vahid Behzadan. Psf-med: Measuring and explaining paraphrase sensitivity in medical vision language models. *arXiv preprint arXiv:2602.21428*, 2026.

[30] T. A. J. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. van den Bosch. Human-centered XAI: developing design principles for explainable AI in clinical practice. *Artificial Intelligence in Medicine*, 2021.

[31] Jordan Sim, Qi Wei Fong, Weimin Huang, and Cher Heng Tan. Machine learning in medicine: what clinicians should know. *Singapore Medical Journal*, 2021.

[32] Christine Sinsky, Rachel Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Judy Westbrook, Michael Tutty, and Mark W. Blanke. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Annals of Internal Medicine*, 165(11):753–760, 2016.

[33] Erico Tjoa and Cuntai Guan. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[34] Angela Zhang, Lei Xing, James Zou, and Joseph C. Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 2022.