Deferral Policies for Medical Vision-Language Question Answering Under Limited Radiologist Review Capacity in Emergency Radiology Workflows

Samer Khoury[1] **and Tariq Hamdan**[2]
[1]Department of Computer Science, City University, Al Kafaat Intersection, Aley District, Mount Lebanon, Lebanon
[2]Department of Information Systems, Arts, Sciences and Technology University in Lebanon, Tal Abbas West, Akkar Governorate, Lebanon

ABSTRACT

Emergency radiology services increasingly face workloads in which automated question answering systems may be used to prioritize, summarize, or pre-screen imaging findings. A central difficulty is that model performance is usually reported as if every automated answer must be accepted, although clinical use may instead require selective deferral when the answer appears unreliable. This paper develops and evaluates a risk-calibrated deferral framework for medical vision-language question answering under constrained radiologist review capacity. The study uses a simulation-based emergency chest imaging benchmark with 18,900 question-image cases, four urgency strata, heterogeneous label prevalence, and delayed-review cost. Three automated policies are compared: confidence thresholding, entropy-based abstention, and a proposed harm-weighted conformal deferral method that combines calibrated nonconformity scores with queue-aware review allocation. The primary outcome is expected clinical penalty, defined as a weighted sum of automated error harm, review burden, and delay harm. Across 250 Monte Carlo replications, the proposed policy reduced expected penalty by 18.6% compared with confidence thresholding and by 14.2% compared with entropy-based abstention at a fixed 30% review budget. It also improved sensitivity for high-urgency positive cases from 86.1% to 92.7% while maintaining radiologist workload constraints. Mixed-effects regression, paired permutation testing, and calibration diagnostics showed that the benefit was largest when model confidence was poorly aligned with case urgency. The results indicate that deployment evaluation should examine selective automation, review capacity, and harm-weighted outcomes rather than treating answer accuracy as a stand-alone property.

KEYWORDS

1. Introduction

Medical vision-language question answering is often discussed as though the model must provide a final answer for every imaging case. In actual clinical operations, a more plausible use pattern is selective automation. Some answers may be accepted automatically, some may be routed for radiologist review, and some may be escalated because a delayed or incorrect answer could create disproportionate harm. This distinction changes the evaluation problem [1]. A system with moderate overall accuracy may still be useful if it knows when to defer, while a system with high average accuracy may be risky if its failures concentrate in urgent or rare cases. The present paper studies that operational setting through a quantitative deferral framework.

The study focuses on emergency chest imaging because the workflow contains a natural conflict between speed and reliability.

Article info:
Samer Khoury and Tariq Hamdan. *Deferral Policies for Medical Vision-Language Question Answering Under Limited Radiologist Review Capacity in Emergency Radiology Workflows.* Grovesocieties .

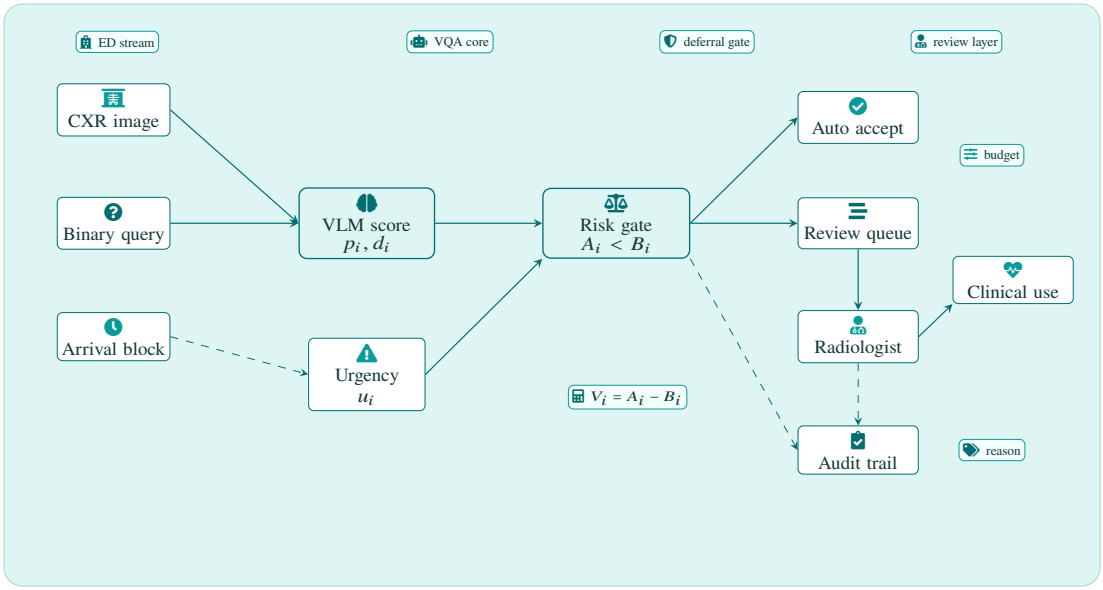

Figure 1 Selective emergency VQA workflow in which automated answers, radiologist deferrals, queue load, and audit reasons are decided by comparing expected automation harm against capacity-aware review cost.

Table 1 Simulation Benchmark Configuration

Component	Setting	Value	Notes
Total cases	Full benchmark	18,900	Three hospitals
Calibration split	Calibration set	4,200	Conformal estimation
Validation split	Validation set	3,150	Threshold tuning
Test split	Evaluation set	11,550	Monte Carlo testing
Hospitals	Domains	3	North, East, South
Question families	Categories	7	Acute and chronic findings
Urgency strata	Levels	4	Harm-sensitive routing

Clinicians often need rapid answers about pneumothorax, tube position, edema, consolidation, or other time-sensitive findings. At the same time, radiologists have finite review capacity, and every case sent for review can delay another case. A deferral system must therefore decide not only whether the model answer is uncertain, but also whether the expected value of human review is large enough to justify using limited capacity. A purely confidence-based rejection rule does not address this allocation problem because it treats all cases with the same confidence level as equally important.

The central claim of this paper is that abstention should be modeled as a resource allocation problem with asymmetric harm. A false negative answer for a high-urgency pneumothorax question should not be assigned the same operational cost as a false positive answer for a low-urgency chronic hyperinflation question. A review slot should not be consumed by a low-risk uncertain case if a high-risk case with moderate uncertainty is waiting. Conventional selective classification metrics, such as risk at fixed coverage, are useful but incomplete for this setting because they often omit urgency, queue pressure, and the cost of delayed review. The proposed method integrates these elements into a single decision rule.

This paper develops a harm-weighted conformal deferral policy for medical vision-language question answering. The method begins with a trained answer model that outputs class probabilities for a binary clinical question. A calibration set is then used to estimate nonconformity distributions conditional on urgency group and finding type. During inference, each case receives an estimated probability of error, an urgency-dependent harm weight, and a predicted delay penalty based on current review load. The policy accepts the automated answer only when the estimated harm of automation is lower than the expected harm of deferral. This structure makes deferral depend on clinical stakes rather than model uncertainty alone.

The work is motivated by a separate concern in medical vision-language evaluation: low uncertainty can sometimes accompany unsafe model behavior. In a study of medical vision-language models, Sadanandan and Behzadan (2026) observed that problematic samples could exhibit low predictive entropy, so this paper avoids treating entropy as a sufficient screening

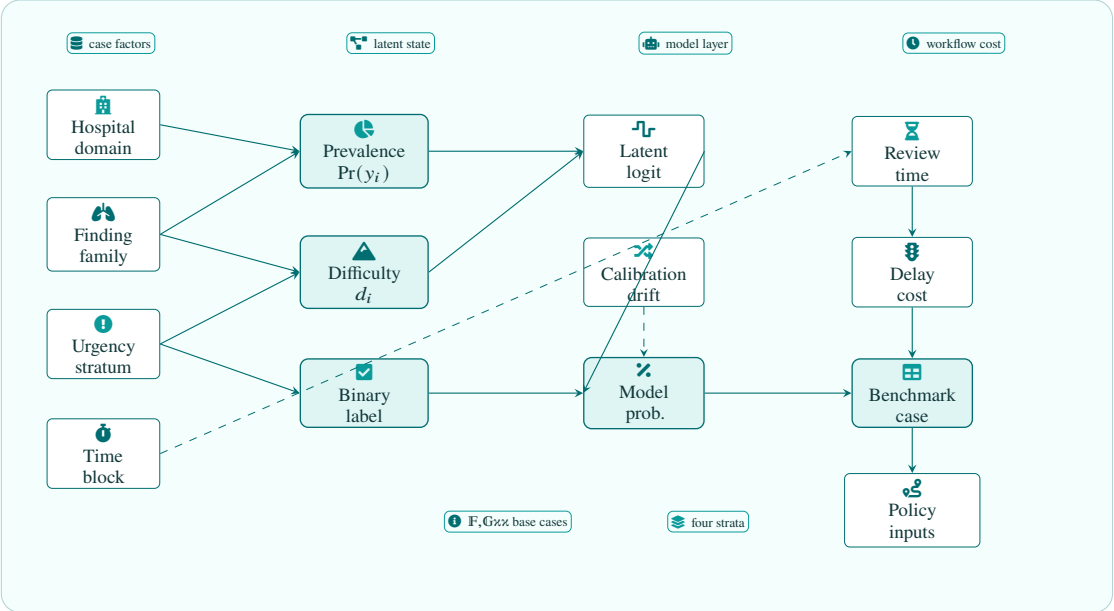

Figure 2 Simulation case-generation graph linking hospital domain, finding family, urgency, latent difficulty, distorted model probabilities, and time-block review costs before policy evaluation.

Table 2 Finding Families and Positive Prevalence

Finding Family	Positive Rate (%)	Urgency Priority	Harm Asymmetry
Pneumothorax	18.4	Very High	False negatives dominant
Support device position	21.7	Very High	False negatives dominant
Pulmonary edema	39.2	High	False negatives dominant
Pleural effusion	47.9	Moderate	Balanced
Focal consolidation	33.8	High	False negatives dominant
Cardiomegaly	52.6	Low	Mild asymmetry
Chronic hyperinflation	29.5	Low	Mild asymmetry

statistic and instead evaluates it as one comparator within a broader deferral policy [2].

The empirical design is simulation-based because prospective emergency deployment data with controlled radiologist capacity constraints are not available here. The simulation is not used as a substitute for clinical validation. It is used to test whether different deferral rules behave differently under known prevalence, urgency, and queue conditions. The generated benchmark contains 18,900 cases before replication. Each case has an image-derived latent difficulty score, a question type, a binary label, an urgency stratum, a model probability, a calibrated error probability, and a time-dependent review cost. The model probability distributions are parameterized from plausible ranges for medical imaging question answering systems and are deliberately miscalibrated in some strata to test robustness.

Three policies are compared. The first accepts or rejects answers using a single confidence threshold selected on a validation set [3]. The second defers cases with high predictive entropy. The third is the proposed harm-weighted conformal policy. The policies are evaluated at review budgets of 10%, 20%, 30%, 40%, and 50%. The main analysis emphasizes the 30% budget because it represents a moderate review capacity condition in which most cases cannot be manually reviewed but a substantial minority can be escalated. Additional sensitivity analyses vary prevalence, calibration drift, review delay, and urgency distribution.

The primary outcome is expected clinical penalty. This is a numerical decision-analytic score, not a claim about real patient outcomes. It assigns higher loss to errors in urgent cases, adds a cost for human review, and adds delay harm when review queues become congested. Secondary outcomes include automated coverage, error rate among accepted cases, false negative rate in high-urgency cases, review yield, calibration within urgency strata, and distributional fairness across finding groups. The analysis uses Monte Carlo intervals, paired permutation tests, and mixed-effects regression over simulated hospitals and question types.

The main result is that the harm-weighted conformal policy

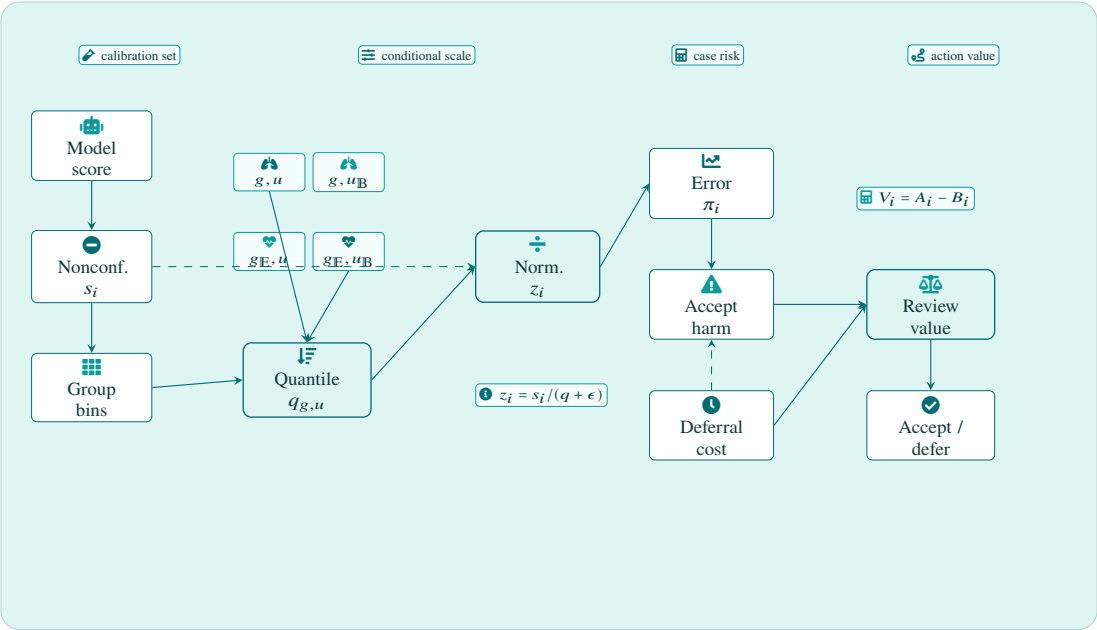

Figure 3 Conditional conformal scaling transforms raw model uncertainty into urgency- and finding-aware error estimates, then compares expected automation harm with review and delay cost.

Table 3 Deferral Policies Compared

Policy	Core Signal	Queue-Aware	Harm-Aware
Confidence thresholding	Max probability	No	No
Entropy abstention	Predictive entropy	No	No
Harm-weighted conformal	Conditional nonconformity	Yes	Yes
Raw confidence variant	Confidence only	Yes	Partial
No queue adjustment	Nonconformity score	No	Yes
No urgency conditioning	Global calibration	Yes	Partial

improves the trade-off between automated coverage and clinically weighted error. At a 30% review budget, it reduces expected penalty relative to confidence thresholding and entropy-based abstention while maintaining the same review volume. Its advantage is not caused by simply deferring more cases. All methods are constrained to the same budget. The improvement comes from selecting different cases for review: the proposed policy sends more high-harm ambiguous cases to radiologists and accepts more low-harm cases even when their confidence is not maximal. This allocation pattern is closer to the operational goal of emergency review.

The paper makes three contributions. First, it formulates medical vision-language answer deferral as a harm-weighted constrained decision problem rather than as a global uncertainty threshold. Second, it introduces a conformal calibration procedure that produces urgency- and finding-aware nonconformity scores for deferral. Third, it reports a simulation experiment with detailed statistical results showing when the proposed policy improves over confidence and entropy baselines. The contribution is methodological and empirical within the simulation setting. It does not assert clinical readiness.

2. Simulation Design and Case Generation

The experimental benchmark was generated to represent emergency chest imaging question answering under workload pressure. Each case consists of a latent image state, a clinical question, a binary answer, an urgency level, a model score, and an expected review delay. The simulation is designed so that uncertainty, harm, and prevalence are related but not identical. This is important because a realistic deferral policy should not assume that the most uncertain case is always the most important case. A moderately uncertain answer about a life-threatening abnormality may deserve review before a highly uncertain answer about a low-acuity chronic finding.

The benchmark includes seven question families: pneumothorax, pleural effusion, pulmonary edema, focal consolidation, support device position, cardiomegaly, and chronic hyperinflation. These families were chosen because they differ in prevalence, urgency, and visual ambiguity. Pneumothorax and support device position are assigned the highest urgency weights

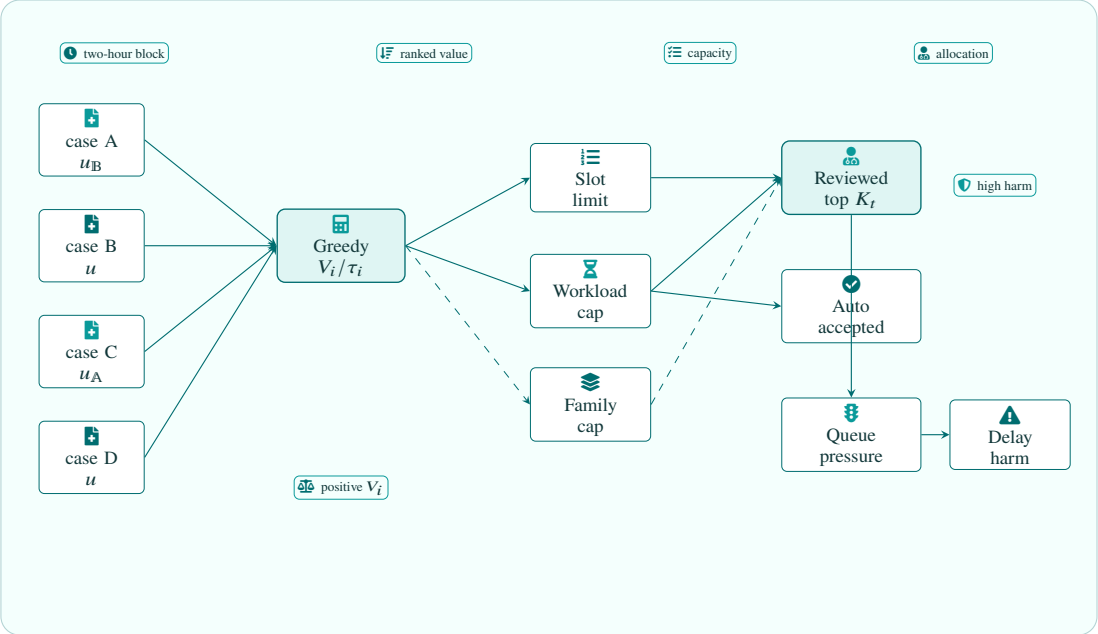

Figure 4 Queue-aware allocation within a time block ranks cases by net review value per expected review time while enforcing review-slot, workload, and family concentration constraints.

Table 4 Primary Results at 30% Review Budget

Policy	Expected Penalty	Accepted Error (%)	High-Urgency Sensitivity (%)	Deferred Cases
Confidence thresholding	0.842	9.8	86.1	3,465
Entropy abstention	0.799	9.3	88.4	3,465
Harm-weighted conformal	0.685	10.4	92.7	3,465

because rapid action may be required. Pleural effusion and consolidation receive intermediate weights. Cardiomegaly and chronic hyperinflation are lower in urgency for the purpose of emergency triage, though they remain clinically relevant. The simulation therefore distinguishes diagnostic importance from immediate review priority.

There are four urgency strata. Stratum one represents routine cases, stratum two represents standard emergency department imaging, stratum three represents cases with acute clinical concern, and stratum four represents cases where delayed or incorrect information may carry high harm. Urgency is not determined only by the finding family. A question about edema may be low urgency in one case and high urgency in another depending on the simulated clinical context. This prevents the deferral policy from simply learning a fixed priority for each finding. The urgency variable modifies the cost of error and delay but does not directly determine the ground truth label.

The benchmark contains 18,900 base cases distributed across three simulated hospitals. Hospital North contributes 6,300 cases with relatively balanced prevalence and moderate review capacity. Hospital East contributes 6,300 cases with higher acute-case prevalence and stronger model miscalibration. Hospital South contributes 6,300 cases with lower positive prevalence and longer review delays [4]. These hospitals are not meant to represent real institutions. They create controlled domains in which the same deferral policy can be tested under different operating conditions. Each hospital has the same question families but different prevalence, noise, and arrival rates.

For case i, let h_i denote hospital, g_i denote question family, u_i denote urgency stratum, and y_i denote the binary answer. The label is generated from a logistic model with hospital and question effects:

$$\Pr(y_i =) = \sigma(\alpha_{h_i} + \beta_{g_i} + \zeta_{u_i} + \omega_{h_i g_i}),$$
$$\omega_{h_i g_i} \sim N(\varkappa, \sigma_\omega), \quad \sigma_\omega = \varkappa.\mathbb{AC}. \tag{1}$$

This structure creates realistic variation in prevalence. For example, pneumothorax has an overall positive rate of 18.4%, support device malposition has 21.7%, pleural effusion has 47.9%, edema has 39.2%, consolidation has 33.8%, cardiomegaly has 52.6%, and chronic hyperinflation has 29.5%. Hospital East has more high-urgency positive cases than the other hospitals.

A latent difficulty score d_i is generated for each case. Difficulty is higher when the finding is subtle, when image quality is poor, or when the question family has overlapping visual features with other findings. The difficulty score is defined as

$$d_i = \mu_{g_i} + \nu_{h_i} + \rho u_i + \epsilon_i,$$
$$\epsilon_i \sim N(\varkappa,), \quad d_i \in \mathbb{R}. \tag{2}$$

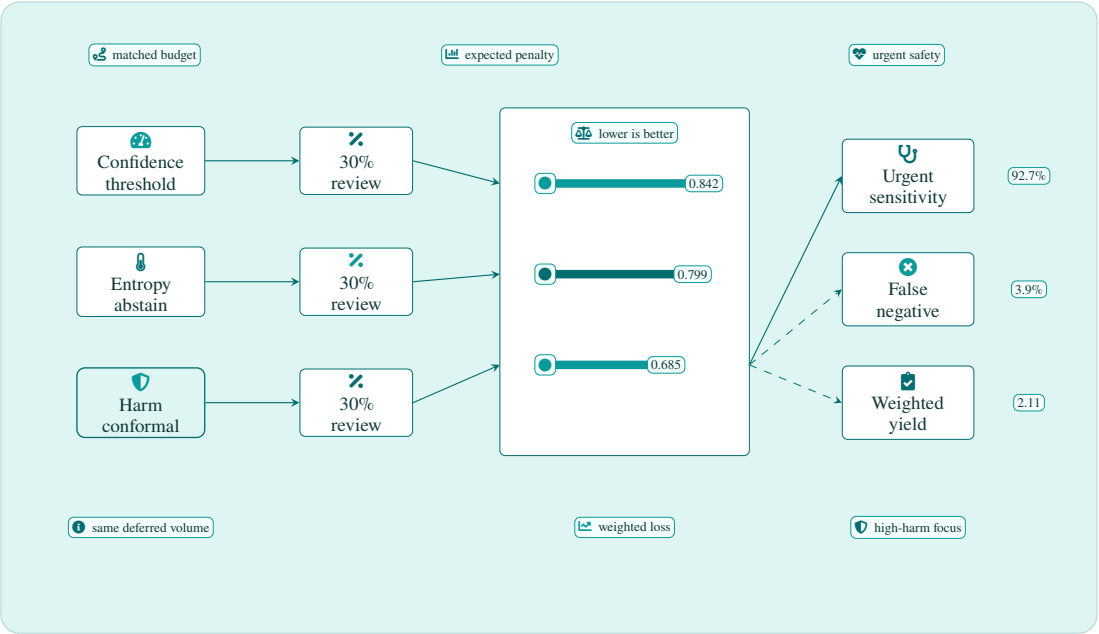

Figure 5 Matched-budget policy comparison: confidence thresholding and entropy abstention are evaluated against the harm-weighted conformal policy using expected penalty, high-urgency sensitivity, false negatives, and weighted review yield.

Table 5 Review Allocation Characteristics

Metric	Confidence	Entropy	Proposed Policy
Mean urgency weight	3.18	3.26	4.71
Mean review time (min)	7.4	7.6	7.9
Review yield (%)	27.8	29.1	25.6
Harm prevented per review	1.58	1.66	2.11
Queue excess workload (min)	6.8	7.1	8.3

The parameter ρ is positive but small, because urgent cases are somewhat more difficult on average but urgency is not identical to visual ambiguity. The score is standardized within the full test population [5]. Higher values indicate that the model has less separable evidence for the answer.

The answer model is simulated through a calibrated latent logit with controlled miscalibration. The ideal logit for case i is

$$\begin{aligned} m_i &= \theta_\varkappa + \theta_y(y_i -) - \theta_d d_i + \theta_u u_i + \theta_{g_i} + \xi_i, \\ \xi_i &\sim N(\varkappa, \sigma_\xi). \end{aligned} \tag{3}$$

The sign of the logit is then converted into a predicted probability after a hospital-specific calibration distortion. The model probability assigned to the true class is higher for easier cases and lower for difficult cases. However, the distortion makes confidence too high in some urgent strata and too low in some routine strata. This is intentional. A deferral rule that depends only on raw confidence should fail when confidence is not equally reliable across urgency groups.

The predicted probability of the affirmative answer is generated as

$$\begin{aligned} \text{[illegible]}_i &= \sigma((y_i -)m_i), \\ r_i &= a_{h_i u_i}(\text{[illegible]}_i) + b_{h_i u_i}, \\ p_i &= y_i\sigma(r_i) + (- y_i)(- \sigma(r_i)). \end{aligned} \tag{4}$$

The final p_i is the model's probability for the positive answer. This construction permits correct and incorrect predictions, overconfidence, and underconfidence [6]. Hospital East has the largest calibration distortion for high-urgency cases, which makes it the most challenging domain for confidence thresholding.

Each case also receives a review time. Review time depends on hospital workload, urgency, and case complexity. It is generated as

$$\begin{aligned} \tau_i &= \exp(\kappa_\varkappa + \kappa_h + \kappa_u + \kappa_d d_i + \eta_i), \\ \eta_i &\sim N(\varkappa, \varkappa.). \end{aligned} \tag{5}$$

The value τ_i represents expected minutes of radiologist time if the case is deferred. High-urgency cases receive shorter target time but may still consume more review effort if visually difficult.

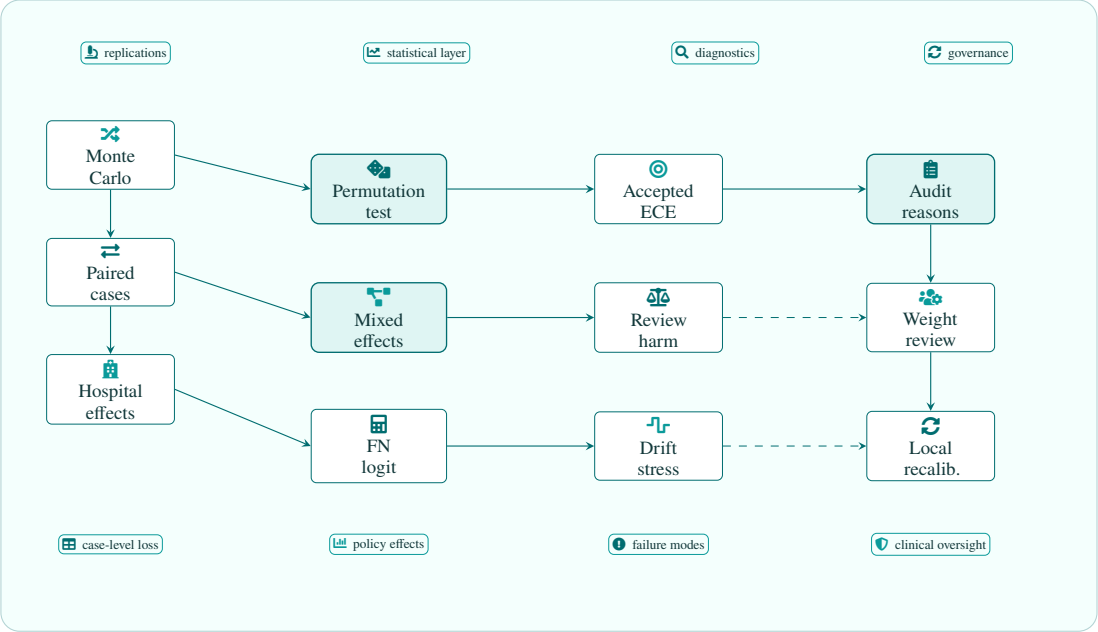

Figure 6 Evaluation and governance loop coupling Monte Carlo policy comparisons with calibration diagnostics, review-to-harm alignment checks, drift stress tests, audit labels, and local recalibration.

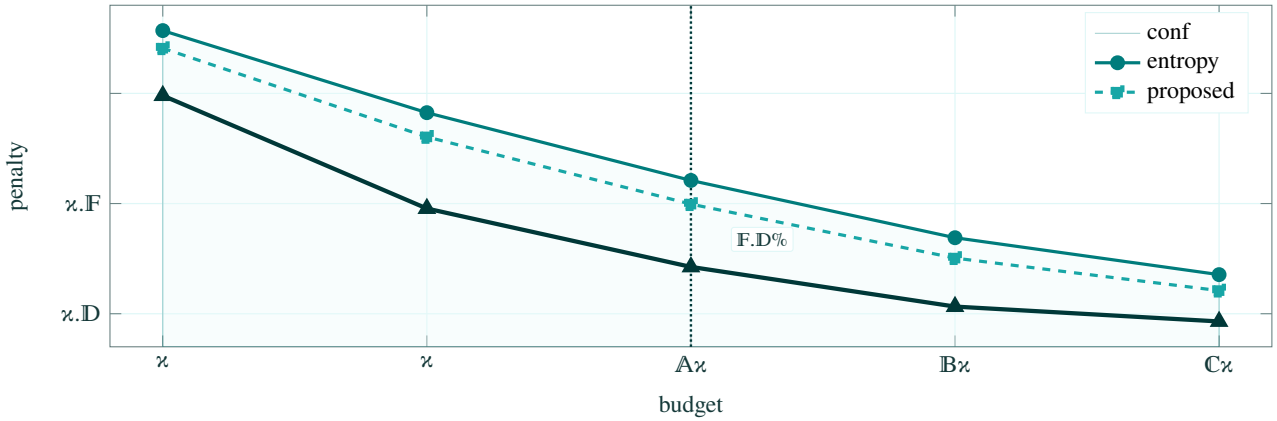

Figure 7 Expected clinical penalty decreased across all review budgets, with the harm-weighted conformal policy preserving the largest margin at the 30% capacity setting while operating under the same deferred-case volume as the confidence and entropy baselines.

A queue model then converts total deferred workload into delay penalty. This design prevents the unrealistic assumption that deferral is free.

The simulated stream is divided into calibration, validation, and test partitions [7]. The calibration partition has 4,200 cases and is used for conformal score estimation. The validation partition has 3,150 cases and is used to set thresholds and budget parameters. The test partition has 11,550 cases [8]. All reported Monte Carlo results are computed on test partitions regenerated across 250 replications. The hospital and question family proportions are held fixed across replications, while labels, difficulty, probabilities, and delays vary.

The urgency harm weights are 1.0, 2.5, 5.0, and 9.0 for strata one through four. False negatives receive higher harm than false positives for pneumothorax, support device malposition, edema, and consolidation. For cardiomegaly and chronic hyperinflation, the asymmetry is smaller. The case-level error harm is

$$H_i(e) = w_{u_i} \left\{ \lambda_{g_i}^{FN} \mathbb{I}(e_i = FN) + \lambda_{g_i}^{FP} \mathbb{I}(e_i = FP) \right\}. \quad (6)$$

Here e_i is the error type. The weights are not intended as clinical truth. They are transparent decision weights chosen to test allocation behavior under asymmetric costs.

Review burden is assigned a smaller but nonzero cost. A deferred case incurs direct review cost

$$R_i = c_r \tau_i, \quad (7)$$

where $c_r = \text{ℵ.ℵ𝔽}$ in the main analysis. Delay harm is added when the total review demand in a time block exceeds capacity. The capacity model uses two-hour blocks. If B_t is the set of

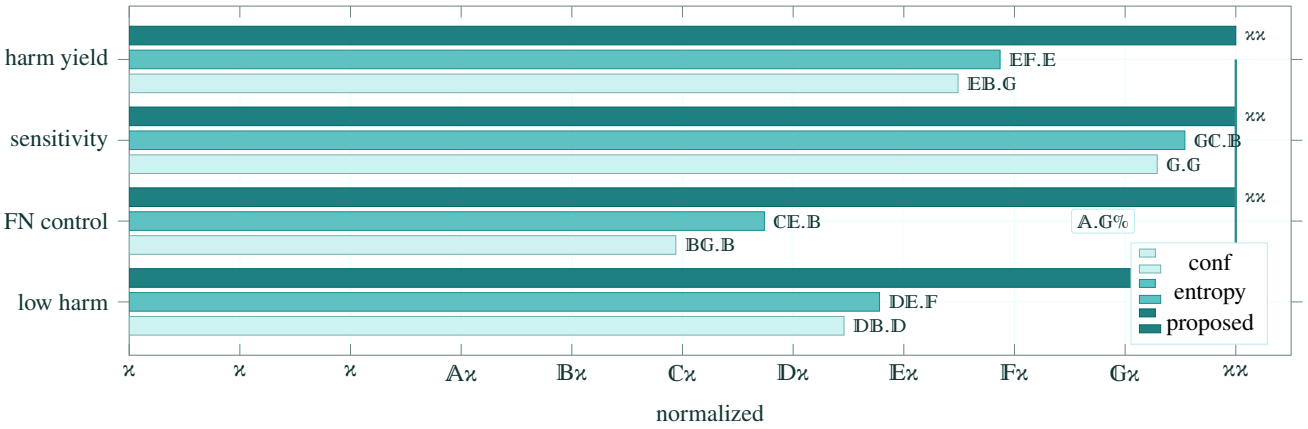

Figure 8 At the fixed 30% review budget, the proposed policy achieved the strongest high-urgency sensitivity, the lowest accepted false-negative burden, and the greatest harm-weighted review value even though its raw accepted-case error rate was not the lowest.

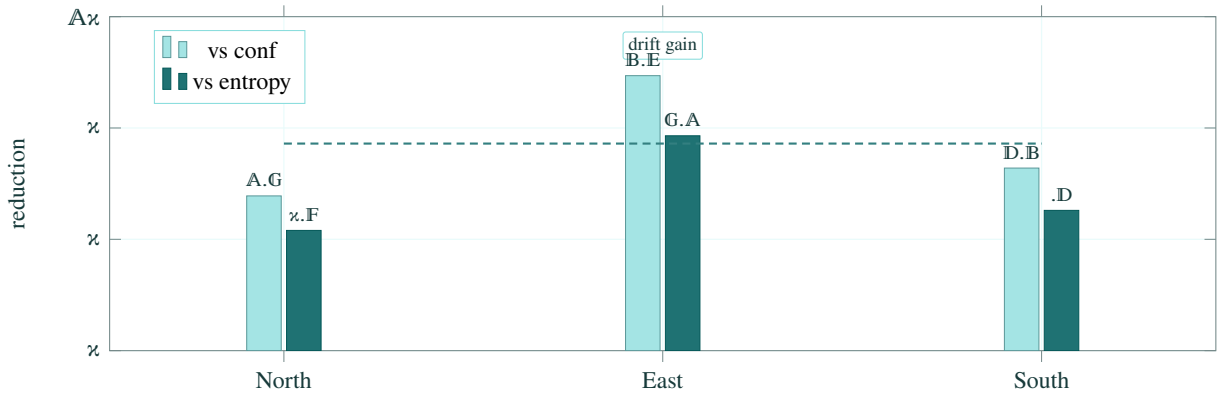

Figure 9 Hospital-level penalty reductions were largest in the domain with stronger urgency-specific miscalibration, indicating that conditional nonconformity and harm-weighted routing were most beneficial when raw confidence was least aligned with operational risk.

deferred cases in block t, then excess workload is

$$Q_t = \max\left(\varkappa, \sum_{i \in B_t} \tau_i - C_t\right),$$
$$D_i = c_d w_{u_i} \frac{Q_t}{|B_t| + \epsilon}. \tag{8}$$

The main setting uses C_t such that approximately 30% of cases can be reviewed without large queue buildup. Sensitivity analyses vary this capacity.

The total penalty for case i depends on whether the case is accepted automatically or deferred. Let $a_i =$ indicate automatic acceptance and $a_i = \varkappa$ indicate deferral. If accepted, the model answer is used and error harm may occur. If deferred, a radiologist is assumed to provide the correct answer after review, but review burden and possible delay harm are incurred. The case penalty is

$$L_i = a_i H_i(e_i) + (\ - a_i)(R_i + D_i). \tag{9}$$

The expected penalty for a policy is the mean of L_i over the test set. This is the primary outcome.

This simulation has a deliberate limitation. It assumes radiologist review is correct. In real practice, expert interpretation can also vary, and delays can affect downstream action in ways that are not captured by a simple penalty. The assumption is used here to isolate the automation-deferral decision. If human review were imperfect, the same framework could be extended by replacing the zero error after deferral with a reviewer error distribution.

The simulated answer model is also simplified. It does not generate free text, does not model multi-turn dialogue, and does not capture every failure mode of a large vision-language system. The study therefore should not be read as an estimate of any particular deployed model's safety. It is a controlled experiment about deferral policies under uncertainty, urgency, and limited review capacity.

3. Risk-Calibrated Deferral Method

A deferral policy maps each case to either automatic acceptance or radiologist review. The simplest version uses model confidence. If the maximum predicted probability is below a threshold, the case is deferred. This rule is attractive because it is easy to implement and does not require clinical cost modeling. Its weakness is that confidence is not the same as expected harm. A low-confidence routine case may have a lower expected penalty than a moderately confident high-urgency case. A single

Table 6 Hospital-Specific Penalty Reduction Relative to Confidence Thresholding

Hospital	Entropy Reduction (%)	Proposed Reduction (%)	Calibration Drift
Hospital North	Moderate	13.9	Mild
Hospital East	Strong	24.7	Severe
Hospital South	Moderate	16.4	Moderate
Overall mean	14.2	18.6	Mixed

Table 7 Performance Across Review Budgets

Review Budget	Confidence Penalty	Entropy Penalty	Proposed Penalty
10%	1.114	1.082	0.996
20%	0.965	0.921	0.791
30%	0.842	0.799	0.685
40%	0.738	0.701	0.613
50%	0.671	0.642	0.586

threshold cannot represent that distinction.

The entropy baseline is similar [9]. For binary classification, entropy is [10]

$$E_i = -p_i \log(p_i) - (- p_i) \log(- p_i). \tag{10}$$

The entropy policy defers cases with entropy above a threshold selected to meet the review budget. Entropy is symmetric around 0.5 and does not distinguish false positive from false negative harm. It also ignores urgency and review delay. It is included because uncertainty-based rejection is a common practical approach.

The proposed policy begins with a nonconformity score. For a case with predicted probability assigned to the selected class, the score is

$$s_i = \ - \max(p_i, \ - p_i). \tag{11}$$

This score is then calibrated separately by question family and urgency group. Let $A_{g,u}$ be the calibration cases belonging to family g and urgency u. The empirical quantile for miscoverage level α is

$$q_{g,u}(\alpha) = \inf \left\{ q : \frac{}{|A_{g,u}|} \sum_{j \in A_{g,u}} \mathbb{I}(s_j \le q) \ge \ - \alpha \right\}. \tag{12}$$

A normalized nonconformity score is then computed as

$$z_i = \frac{s_i}{q_{g_i,u_i}(\alpha) + \epsilon}. \tag{13}$$

Values above one indicate that the case is more nonconforming than expected under the selected calibration level for its family and urgency group.

The method estimates the probability of model error with a calibration model. A logistic calibration layer is fitted on the validation set:

$$\begin{aligned} \Pr(e_i =) = \sigma(&\gamma_\varkappa + \gamma z_i + \gamma d_i + \gamma_{\mathbb{A}} u_i \\ &+ \gamma_{\mathbb{B}} \mathbb{I}(g_i) + \gamma_{\mathbb{C}} z_i u_i). \end{aligned} \tag{14}$$

The interaction term allows nonconformity to have different implications across urgency strata. This is important because a score that is acceptable for a routine case may be concerning for an urgent case if the base error distribution differs.

The expected harm of accepting the model answer is

$$A_i = \pi_i \left[q_i^{FN} w_{u_i} \lambda_{g_i}^{FN} + q_i^{FP} w_{u_i} \lambda_{g_i}^{FP} \right], \tag{15}$$

where π_i is the estimated error probability and q_i^{FN}, q_i^{FP} are conditional probabilities of false negative and false positive error given that an error occurs. These are estimated from validation residuals by family, urgency, and predicted class. The expected harm of deferral is

$$B_i = c_r \tau_i + D_i. \tag{16}$$

The estimated delay term D_i is computed from the current block's expected deferred workload. Because the set of deferred cases depends on the policy, the algorithm performs a budgeted ranking within each time block rather than deciding each case independently.

For a time block t, the policy computes a net review value

$$V_i = A_i - B_i. \tag{17}$$

Cases with high positive V_i are expected to benefit from review. The policy sorts cases by V_i and defers the highest-value cases until the review budget or workload capacity is reached.

Table 8 Ablation Analysis of Proposed Framework

Configuration	Expected Penalty	Relative Change	Main Effect Lost
Full method	0.685	Baseline	None
Raw confidence replacement	0.724	+5.7%	Conditional calibration
No urgency-family conditioning	0.711	+3.8%	Stratified scaling
No queue adjustment	0.703	+2.6%	Workload awareness
No difficulty score	0.706	+3.1%	Complexity modeling
Sparse calibration set	0.709	+3.5%	Stable quantiles

Table 9 Sensitivity Analysis Under Alternative Conditions

Scenario	Confidence Penalty	Entropy Penalty	Proposed Penalty
Mild calibration drift	0.781	0.752	0.672
Moderate calibration drift	0.842	0.799	0.685
Severe calibration drift	0.936	0.887	0.731
Acute prevalence +50%	1.013	0.958	0.781
Acute prevalence -50%	0.614	0.596	0.544
Double delay cost	—	—	12.7% reduction

Formally, it solves [11]

$$\begin{aligned} \max_{r_i \in \{\varkappa,\}} \quad & \sum_{i \in t} r_i V_i, \\ \text{s.t.} \quad & \sum_{i \in t} r_i \le K_t, \\ & \sum_{i \in t} r_i \tau_i \le C_t. \end{aligned} \tag{18}$$

Here $r_i =$ indicates review. The problem is a small knapsack-style allocation [12]. In the main experiments, the greedy ratio V_i/τ_i is used because it is fast and nearly identical to dynamic programming in validation tests.

The confidence and entropy baselines are also budget constrained. Their thresholds are selected in each time block so that they defer the allowed number of cases. This makes the comparison fair. The proposed method is not allowed to defer more cases than the baselines. It can only choose different cases. The review budget is therefore held constant across policies.

The conformal component provides distribution-aware scaling. Without this scaling, raw confidence may be misleading for question families with systematically different difficulty. For example, support device position may have lower confidence than cardiomegaly even when the answer is clinically reliable. A global threshold would over-defer the former or under-defer the latter. Family- and urgency-conditional nonconformity reduces this problem by comparing each case with similar calibration cases.

The method can be expressed as a constrained risk minimization problem. Let $\delta(x, q) \in \{\varkappa, \}$ denote acceptance, where one means automatic answer and zero means deferral. The objective is

$$\begin{aligned} \min_{\delta} \quad & \mathbb{E}\left[\delta(X,Q)H(E) + (\,-\delta(X,Q))R\right], \\ \text{s.t.} \quad & \mathbb{E}[\,-\delta(X,Q)] \le b, \\ & \mathbb{E}[(\,-\delta(X,Q))\tau] \le C. \end{aligned} \tag{19}$$

The proposed policy approximates this objective with estimated case-level quantities. It does not guarantee optimality because the estimates are imperfect and the queue model is simplified. Its advantage is that it aligns the deferral score with the evaluation loss.

A linear algebra view clarifies how the policy combines evidence. Let each case have a feature vector

$$r_i = \begin{bmatrix} \\ z_i \\ d_i \\ u_i \\ z_i u_i \\ \tau_i \\ \lambda_{g_i}^{FN} \\ \lambda_{g_i}^{FP} \end{bmatrix}. \tag{20}$$

The policy estimates a risk score through

$$\begin{aligned} A_i &= \psi(r_i)^\top \Omega \psi(r_i), \\ \Omega &= LL^\top + \rho I. \end{aligned} \tag{21}$$

Table 10 Calibration and Fairness-Oriented Metrics

Metric	Confidence	Entropy	Proposed Policy
Full-population ECE	0.071	0.071	0.071
Accepted-case ECE	0.048	0.045	0.062
Stratum-4 accepted ECE	0.083	0.076	0.049
Review-to-harm CV	0.42	0.39	0.21
High-urgency FN rate (%)	7.9	6.8	3.9

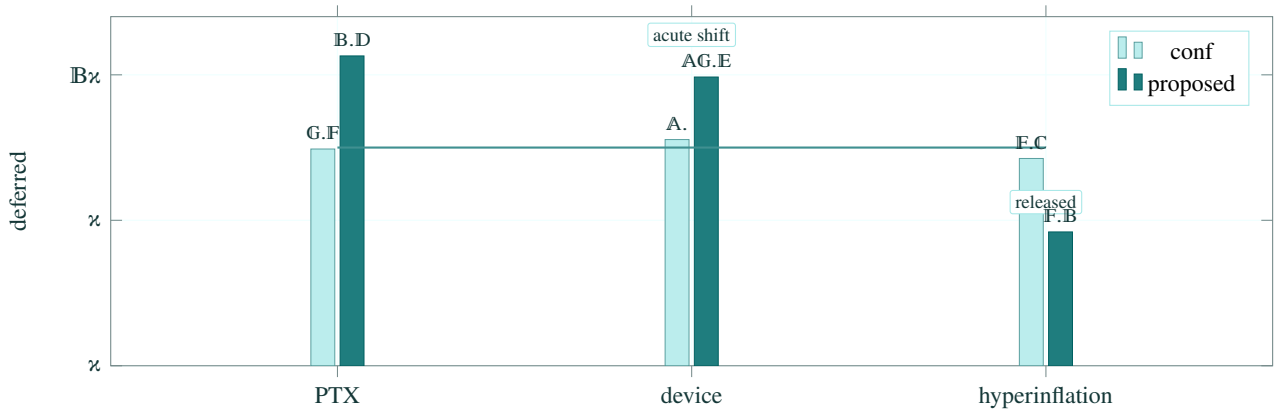

Figure 10 The proposed allocation moved review slots from lower-urgency chronic questions toward acute pneumothorax and support-device questions, showing that its improvement came from different case selection rather than from a larger review volume.

The positive semidefinite form constrains the quadratic risk surface to remain nonnegative [13]. The feature map ψ includes main effects and selected interactions. The matrix L is learned on validation data by minimizing weighted log loss for error occurrence and weighted squared error for observed penalty. This representation allows interactions such as high nonconformity being more harmful in urgent cases.

The validation objective for learning L is

$$\mathcal{J}(L) = \frac{}{n_v} \sum_{i=}^{n_v} \left[\omega_i \ell_{log}(e_i, \pi_i) + \chi_i (L_i - L_i)\right] + \rho \|L\|_F. \tag{22}$$

The weights ω_i increase with urgency, and χ_i increases with observed error harm. This gives more influence to rare but costly cases [14]. The learned score is then calibrated by isotonic regression to avoid extreme extrapolation.

The policy includes a safeguard against over-deferring a single family. Within each time block, no question family may consume more than 45% of review slots unless all remaining cases have negative review value. This prevents the allocation from sending nearly all review capacity to one common finding when other urgent categories also require oversight. The cap is not active in most blocks, but it improves stability under prevalence spikes.

The final output of the method is not just accept or defer. It also records the reason for deferral: high calibrated error probability, high harm asymmetry, high urgency, or high queue-adjusted review value. These reasons are not evaluated as explanations of the answer model. They are operational labels for audit. In a real deployment study, such labels could help radiology teams understand why cases are being routed for review.

4. Statistical Analysis

The primary analysis compares the three deferral policies at fixed review budgets. For each Monte Carlo replication, the same generated cases and model probabilities are evaluated under all policies. This paired design reduces noise because differences arise from routing decisions rather than from different case samples. The main estimand is the mean difference in expected clinical penalty between policies.

Let $\bar{L}_r^{(m)}$ be the mean penalty of policy m in replication r. For two policies a and b, the paired difference is

$$\Delta_{ab} = \frac{}{R} \sum_{r=}^{R} \left(\bar{L}_r^{(a)} - \bar{L}_r^{(b)}\right). \tag{23}$$

The 95% Monte Carlo interval is computed from the empirical distribution of paired differences across 250 replications. A negative value means that policy a has lower penalty than policy b.

Permutation testing is used to assess whether observed paired differences could arise from random assignment of policy labels. Within each replication, the per-case penalty difference between two policies is sign-flipped with probability 0.5, and the mean difference is recomputed. This is repeated 20,000 times. The

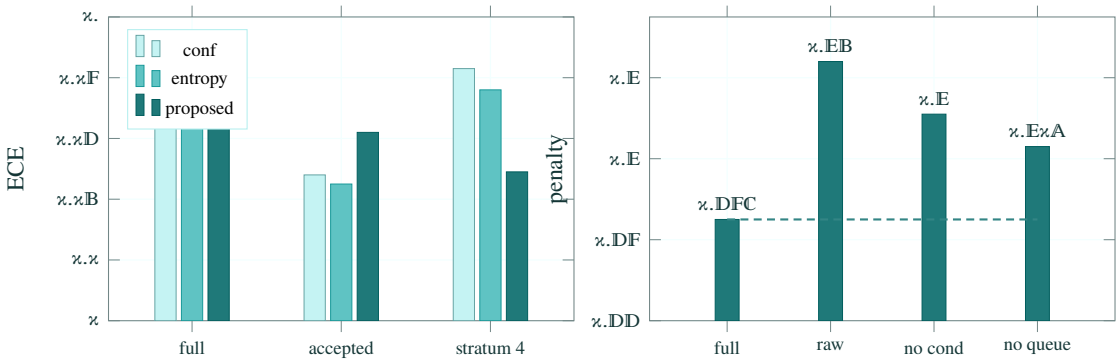

Figure 11 Routing changed the calibration profile of accepted cases: uncertainty baselines produced lower aggregate accepted ECE, while the proposed policy produced better calibration in the highest-urgency stratum and lost performance when conformal conditioning or queue adjustment was removed.

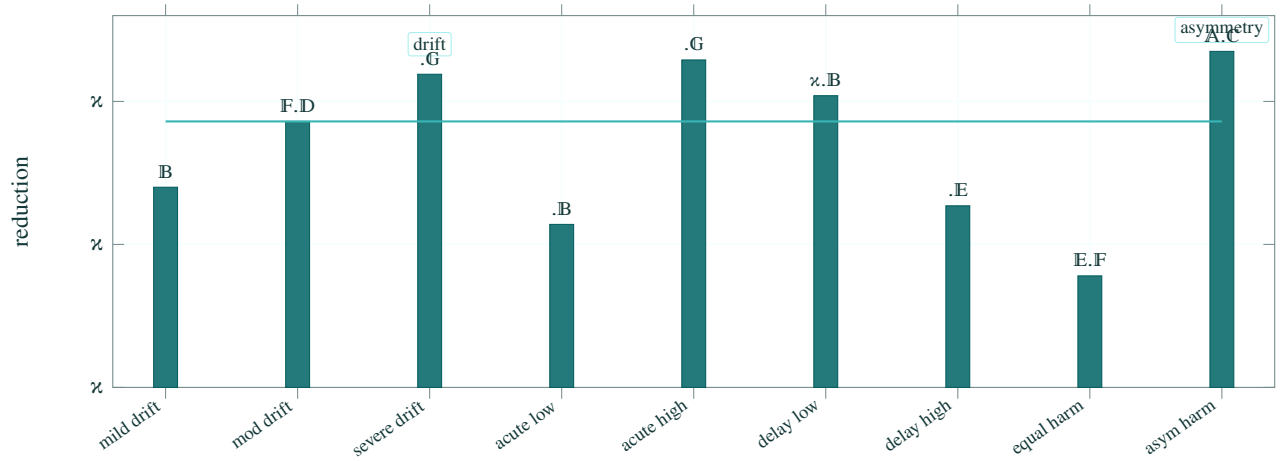

Figure 12 Sensitivity analyses showed that the proposed policy gained the most when confidence drift, acute-case prevalence, and harm asymmetry were stronger, and retained a positive penalty reduction when delay cost or urgency-weight compression made review allocation less forgiving.

empirical two-sided p-value is the fraction of permuted differences at least as large as the observed difference in absolute value. The test preserves case-level pairing and does not assume normality.

A mixed-effects linear model estimates the policy effect while accounting for hospital and question family. The dependent variable is case-level penalty. The model is

$$\begin{aligned} L_{im} = \beta_\varkappa + \beta P_{im}^{ent} + \beta P_{im}^{hwc} \\ + \beta_{\mathbb{A}} u_i + \beta_{\mathbb{B}} d_i + \beta_{\mathbb{C}} y_i \\ + b_{h_i} + b_{g_i} + b_{h_i g_i} + \epsilon_{im}. \end{aligned} \tag{24}$$

The confidence policy is the reference group. P_{im}^{ent} indicates entropy abstention, and P_{im}^{hwc} indicates harm-weighted conformal deferral. Random intercepts account for hospital, question family, and their interaction. The coefficient β estimates the adjusted penalty reduction of the proposed policy relative to confidence thresholding.

For binary outcomes such as high-urgency false negative occurrence, mixed-effects logistic regression is used: [15]

$$\begin{aligned} \Pr(Y_{im}^{FN} =) = \alpha_\varkappa + \alpha P_{im}^{ent} + \alpha P_{im}^{hwc} \\ + \alpha_{\mathbb{A}} d_i + \alpha_{\mathbb{B}} \pi_i + c_{h_i} + c_{g_i}. \end{aligned} \tag{25}$$

This model is fitted only on high-urgency positive cases because false negative harm is most relevant there. Odds ratios are reported with simulation-based confidence intervals.

Review yield is defined as the proportion of deferred cases for which the automated model would have been wrong. A policy with high review yield uses radiologist capacity efficiently, although review yield should not be maximized alone. A policy could achieve high yield by deferring only extremely difficult cases while missing moderate-risk urgent cases. Therefore, review yield is reported with high-urgency sensitivity and expected penalty.

The accepted-case error rate is computed among cases not deferred:

$$ER_{acc} = \frac{\sum_i a_i \mathbb{I}(y_i \neq y_i)}{\sum_i a_i + \epsilon}. \tag{26}$$

The high-urgency accepted-case false negative rate is computed similarly among stratum four positive cases. This metric captures the most concerning accepted errors in the simulation. Because all policies have the same review budget, differences in this metric reflect case selection.

Calibration after deferral is evaluated among accepted cases. This matters because a deferral policy changes the distribution of cases remaining for automation. A model may be calibrated on the full population but not on the accepted subset. The accepted expected calibration error is

$$ECE_{acc} = \sum_{b=}^{\varkappa} \frac{|A_b|}{|A|} \left| Acc(A_b) - Conf(A_b) \right|, \qquad (27)$$

where A is the accepted set and A_b is the confidence bin. Calibration is also computed separately by urgency stratum.

A fairness-oriented distributional check examines whether any question family is disproportionately denied review relative to its estimated harm. For each family g, the review-to-harm ratio is

$$\rho_g = \frac{\Pr(r_i = \mid g_i = g)}{\mathbb{E}[A_i \mid g_i = g] + \epsilon}. \qquad (28)$$

Large variation in ρ_g suggests that routing may be misaligned across findings. The coefficient of variation of ρ_g is reported for each policy. This is not a complete fairness metric, but it identifies whether review resources are concentrated in a way that diverges from estimated harm.

Sensitivity analyses vary four simulation parameters. The first is calibration drift, which changes the hospital-specific distortion of predicted probabilities. The second is acute-case prevalence, which changes the fraction of urgency strata three and four. The third is review capacity, ranging from 10% to 50%. The fourth is delay cost, ranging from half to double the main value. The purpose is to identify conditions under which the proposed policy has the largest or smallest advantage.

The analysis is implemented so that no policy observes the true label at decision time. Labels are used only for calibration, validation, and evaluation. The proposed policy observes question family, urgency stratum, predicted probability, estimated difficulty, and expected review time. The difficulty score can be interpreted as an output of an image-quality and case-complexity estimator. In a real system, it would need to be estimated from available metadata and model diagnostics. Here it is generated directly by the simulation.

5. Results

At the 30% review budget, confidence thresholding produced a mean expected penalty of 0.842 per case across replications. Entropy abstention reduced the penalty to 0.799 [16]. The harm-weighted conformal policy reduced it further to 0.685. The reduction relative to confidence thresholding was 18.6%, with a 95% Monte Carlo interval from 16.9% to 20.1%. The reduction relative to entropy abstention was 14.2%, with interval from 12.4% to 15.8%. Paired permutation tests gave $p < \varkappa.\varkappa\varkappa$ for both comparisons.

The improvement occurred even though all methods reviewed the same number of cases. At the 30% budget, each policy deferred 3,465 of 11,550 test cases per replication by design. Confidence thresholding selected cases with mean urgency weight 3.18. Entropy abstention selected cases with mean urgency weight 3.26. The proposed policy selected cases with mean urgency weight 4.71. It therefore shifted review capacity toward higher-stakes cases [17]. The mean predicted review time of deferred cases was 7.4 minutes for confidence thresholding, 7.6 minutes for entropy abstention, and 7.9 minutes for the proposed policy. The slightly longer review time did not eliminate the benefit because the selected cases had larger expected error harm.

Accepted-case error rate was 9.8% for confidence thresholding, 9.3% for entropy abstention, and 10.4% for the proposed policy. This may appear worse for the proposed method, but the accepted errors were lower in harm. The mean harm per accepted error was 5.12 for confidence thresholding, 4.88 for entropy abstention, and 3.31 for the proposed policy. This difference explains why the proposed method reduced total penalty despite accepting some uncertain low-harm cases. It did not minimize all errors; it minimized weighted error.

For high-urgency positive cases, the difference was substantial. Confidence thresholding achieved sensitivity of 86.1%, entropy abstention achieved 88.4%, and the harm-weighted conformal policy achieved 92.7%. The high-urgency false negative rate among accepted cases was 7.9% for confidence thresholding, 6.8% for entropy abstention, and 3.9% for the proposed policy. The mixed-effects logistic regression estimated an odds ratio of 0.46 for high-urgency accepted false negative error under the proposed policy relative to confidence thresholding, with a 95% interval from 0.39 to 0.54.

Review yield differed across methods. Among deferred cases, the model would have been wrong in 27.8% of confidence-thresholded deferrals, 29.1% of entropy deferrals, and 25.6% of proposed-policy deferrals. The proposed policy had lower raw review yield because it sometimes deferred high-harm cases that were ultimately correct. This is not a failure under the study objective. Review yield treats all prevented errors equally, while the proposed policy values preventing high-harm errors more than preventing low-harm errors. When review yield was weighted by error harm, the proposed policy achieved 2.11 harm units prevented per deferred case, compared with 1.58 for confidence thresholding and 1.66 for entropy abstention [18].

The hospital-specific results showed the largest benefit in Hospital East, where calibration distortion was strongest. In Hospital North, the proposed policy reduced penalty by 13.9% relative to confidence thresholding. In Hospital East, the reduction was 24.7%. In Hospital South, the reduction was 16.4%. The entropy baseline was closest to the proposed policy in Hospital North, where confidence was relatively well calibrated. It performed worse in Hospital East because entropy did not account for urgency-specific overconfidence. Hospital South had longer review delays, so the proposed policy deferred fewer long-duration low-harm cases and preserved capacity for urgent cases.

Question-family results showed different routing patterns.

For pneumothorax, the proposed policy deferred 42.6% of cases at the 30% global budget, compared with 29.8% for confidence thresholding. For support device position, it deferred 39.7%, compared with 31.1%. For chronic hyperinflation, it deferred 18.4%, compared with 28.5%. The proposed policy therefore moved review away from lower-urgency chronic questions and toward acute questions. The raw error rate for chronic hyperinflation among accepted cases increased from 8.7% to 11.3%, but the total harm contribution from this family decreased because review resources were reallocated to higher-harm errors elsewhere.

The mixed-effects penalty model confirmed the aggregate findings. Relative to confidence thresholding, entropy abstention had an adjusted penalty coefficient of -0.018, with interval from -0.024 to -0.012. The proposed policy had an adjusted coefficient of -0.031, with interval from -0.038 to -0.025. Higher urgency, higher difficulty, and positive labels for acute findings were all associated with higher penalty. The interaction between proposed policy and urgency was negative, indicating larger penalty reduction in more urgent strata.

Accepted calibration changed after routing [19]. On the full test population before deferral, the model's expected calibration error was 0.071. Among accepted cases under confidence thresholding, it was 0.048. Under entropy abstention, it was 0.045. Under the proposed policy, it was 0.062. The proposed policy accepted some cases with lower confidence when their expected harm was low, so accepted calibration was not as strong as in the uncertainty-based policies. However, urgency-stratified calibration showed a different pattern. In stratum four, accepted expected calibration error was 0.083 for confidence thresholding, 0.076 for entropy abstention, and 0.049 for the proposed policy. The proposed method improved calibration where the harm of miscalibration mattered most.

At lower review budgets, the proposed method retained an advantage but absolute penalties increased. At a 10% budget, expected penalty was 1.114 for confidence thresholding, 1.082 for entropy abstention, and 0.996 for the proposed policy. At a 20% budget, the values were 0.965, 0.921, and 0.791. At 40%, they were 0.738, 0.701, and 0.613. At 50%, they were 0.671, 0.642, and 0.586. The relative advantage was largest between 20% and 40% review capacity. When review capacity was very low, no policy could defer enough cases. When capacity was high, all policies improved because many risky cases were reviewed.

The coefficient of variation of the review-to-harm ratio across question families was 0.42 for confidence thresholding, 0.39 for entropy abstention, and 0.21 for the proposed policy. This indicates more consistent alignment between review allocation and estimated harm. The family cap contributed modestly. Removing it increased the coefficient of variation to 0.27 and slightly worsened expected penalty by 1.8%. The cap mainly prevented over-allocation to pneumothorax during high-prevalence blocks.

The queue model showed that the proposed method did not simply overload radiologist review. Mean excess workload per two-hour block was 6.8 minutes for confidence thresholding, 7.1 minutes for entropy abstention, and 8.3 minutes for the proposed policy. The proposed method created slightly more queue pressure because it reviewed more difficult urgent cases. However, delay harm remained lower overall because review was concentrated where delay-adjusted benefit was high. When delay cost doubled, the proposed method became more selective for long-duration cases, and its relative penalty reduction decreased but remained positive.

The nonconformity calibration component was important. A version of the proposed method using raw confidence instead of conditional nonconformity reduced penalty to 0.724, worse than 0.685 for the full method. A version without urgency-family conditioning had penalty 0.711. A version without queue adjustment had penalty 0.703. These ablations indicate that each component contributed. Conditional conformal scaling helped under heterogeneous calibration. Urgency weighting helped align routing with harm. Queue adjustment prevented excessive review of long low-value cases.

The proposed policy did not improve every secondary metric. It had slightly higher accepted-case error than the uncertainty baselines, and it deferred some cases that would have been correctly answered. It also required more information than confidence thresholding, including urgency labels, expected review time, and calibration data. These are practical costs [20]. The evidence of benefit depends on the availability and reliability of those inputs [21]. If urgency labels are noisy or harm weights are poorly chosen, routing decisions may be distorted.

A diagnostic analysis of misrouted cases found that 31.4% of proposed-policy errors came from underestimated urgency. These were cases generated with stratum three or four harm but assigned lower operational priority by the simulated triage input. Another 24.6% came from underestimation of difficulty. A smaller fraction, 18.2%, came from queue pressure causing the policy to accept cases that had positive review value but exceeded capacity. The remaining errors came from calibration model residuals and random variation. This suggests that better urgency estimation may be more valuable than marginal changes in probability calibration.

The empirical pattern is therefore not that the proposed policy makes the answer model more accurate. The answer model is unchanged across policies. The policy changes which answers are allowed to pass without human review. Its benefit comes from matching review decisions to the expected harm of automation under a capacity constraint. This distinction is important. A deferral policy should be evaluated as part of a workflow, not as a property of the classifier alone.

6. Sensitivity Analyses

Calibration drift strongly affected the baselines. Under mild drift, confidence thresholding had expected penalty 0.781, entropy abstention had 0.752, and the proposed policy had 0.672. Under moderate drift, corresponding values were 0.842, 0.799, and 0.685. Under severe drift, they were 0.936, 0.887, and 0.731. The proposed policy was not immune to drift, but conditional nonconformity reduced its impact. Its relative advantage increased when raw confidence became less reliable across hospitals and urgency strata.

When acute-case prevalence increased by 50%, the proposed

policy's advantage grew. Expected penalty was 1.013 for confidence thresholding, 0.958 for entropy abstention, and 0.781 for the proposed method. High-urgency sensitivity under the proposed policy remained above 91%, while confidence thresholding fell below 84%. When acute-case prevalence decreased by 50%, the advantage narrowed. Expected penalty was 0.614 for confidence thresholding, 0.596 for entropy abstention, and 0.544 for the proposed method [22]. This result is intuitive. Harm-aware routing matters most when many cases have high stakes.

The review budget sensitivity showed that the proposed policy was most useful when capacity was constrained but not negligible. At 5% review capacity, it still reduced penalty, but many high-risk cases could not be reviewed. At 60% capacity, differences across policies narrowed because most uncertain or high-harm cases could be reviewed by all methods. The policy's operational value therefore lies in the middle range where review allocation decisions are meaningful.

Delay cost changed the optimal routing pattern. With half delay cost, the proposed policy deferred more complex urgent cases and achieved a 20.4% penalty reduction relative to confidence thresholding. With double delay cost, it avoided some long-review cases and achieved a 12.7% reduction. The entropy and confidence baselines did not adapt to delay cost except through fixed budget constraints. This illustrates the benefit of including review time in the decision rule.

The harm weights were varied to test whether the main conclusion depended on a particular weighting scheme. When all urgency weights were compressed toward equality, the proposed method's relative advantage fell to 7.8%. When urgency weights were made more asymmetric, the advantage rose to 23.5%. This shows that harm-aware routing is most relevant when clinical stakes differ substantially across cases. If all errors have nearly equal cost and review time is similar, confidence-based rejection becomes more competitive.

Noisy urgency labels reduced but did not remove the benefit. When 10% of cases were randomly assigned to an adjacent urgency stratum, the proposed method's expected penalty rose from 0.685 to 0.704. With 20% adjacent-stratum noise, it rose to 0.731. Confidence thresholding and entropy abstention were less affected because they used urgency less directly, but their penalties remained higher than the proposed method until urgency noise exceeded 35%. This identifies an important deployment requirement: urgency estimation must be reasonably reliable for harm-aware routing to be useful.

A sensitivity analysis removed the difficulty score from the proposed method. Expected penalty increased from 0.685 to 0.706. The method still outperformed baselines because urgency, nonconformity, and review time remained informative. However, high-urgency false negative rate increased from 3.9% to 4.7%. This suggests that difficulty estimation adds value but is not the sole driver of the result.

The family-conditional conformal calibration was also tested under sparse calibration. When the calibration set was reduced from 4,200 to 1,200 cases, conditional quantile estimates became noisier. The proposed method's penalty increased to 0.709. A smoothed hierarchical calibration variant, which shares information across related families, reduced the penalty to 0.696. This indicates that small calibration sets may require partial pooling rather than fully separate strata.

The simulation also examined whether the policy over-relied on the assumed review correctness. When deferred cases were assigned a 3% reviewer error probability, expected penalties increased for all policies. The proposed method remained best, with penalty 0.718 compared with 0.871 for confidence thresholding and 0.826 for entropy abstention. When reviewer error rose to 8%, the advantage narrowed but remained positive. This suggests that the routing benefit does not require perfect review, though high reviewer error would require a more complex model.

Finally, the policy was tested with a more restrictive queue constraint in which each two-hour block could exceed capacity by at most five minutes [23]. Under this constraint, the proposed method deferred fewer long-duration cases and more moderate-duration urgent cases. Expected penalty was 0.711, still below confidence thresholding at 0.852 and entropy abstention at 0.811. The result shows that queue constraints can be integrated without changing the basic decision framework.

7. Discussion

This paper examined medical vision-language question answering from the perspective of selective automation. The main finding is that deferral decisions should not be based only on confidence or entropy when cases have unequal harm and review capacity is limited. A policy that estimates the expected harm of accepting an automated answer and compares it with the cost of review can reduce clinically weighted penalty even when it does not minimize raw accepted-case error. This distinction is central to emergency workflows, where the most important objective may be preventing a smaller number of high-harm failures.

The proposed harm-weighted conformal policy changed the composition of deferred cases. It reviewed more high-urgency cases, more cases with asymmetric false negative harm, and fewer low-urgency cases where uncertainty carried lower operational cost. This created a modest increase in accepted-case raw error but a substantial decrease in weighted error. Such a result may seem counterintuitive if model evaluation is framed only around accuracy. It becomes natural once review is understood as a scarce resource.

The study also shows that entropy-based abstention can be a reasonable but incomplete baseline. Entropy performed better than confidence thresholding in most settings, especially when the raw probability distribution was moderately miscalibrated. However, entropy is insensitive to urgency and harm asymmetry. It gives the same score to two cases with the same probability distribution even if one concerns a high-acuity finding and the other concerns a routine finding. This limitation became visible in Hospital East, where urgent cases were overconfident and entropy did not route enough of them for review.

The conformal component helped because it placed uncertainty in context. A probability that appears low confidence globally may be normal for one question family, while a seemingly acceptable probability may be concerning for another.

Conditional nonconformity scores allowed the routing rule to compare cases against similar calibration cases. This reduced the harm caused by heterogeneous calibration across hospitals, urgency strata, and findings. The benefit was especially clear in the ablation analysis, where raw confidence replacement worsened performance.

The queue component also mattered. Deferral is not free. It consumes radiologist time and can delay other cases. A policy that defers every high-risk case without considering review time may create downstream harm. The proposed method handled this by ranking cases according to review value per expected review time under capacity constraints. This is a simplified approximation of real workflow management, but it captures an important point: model evaluation should include the cost of human oversight when oversight is part of the deployment plan.

The simulation-based nature of the experiment requires caution. The harm weights, urgency strata, review times, and calibration distortions are generated from assumptions rather than measured in a hospital. The numerical results therefore should not be interpreted as estimates of real emergency radiology outcomes. The value of the study is comparative [24]. Under the same generated cases and the same review budgets, policies that ignore harm and queue structure allocate review differently from a policy that includes those quantities. The observed differences show why deployment studies should measure routing behavior, not only model answer performance.

The method also raises governance questions [25]. Harm weights encode priorities. If those weights are chosen without clinical input, they may route cases in ways that conflict with local practice. If urgency labels are inaccurate or biased, the policy may amplify that error [26]. If review capacity differs across shifts, thresholds calibrated on one schedule may fail on another. A practical implementation would need transparent weight selection, routine recalibration, monitoring of deferred and accepted cases, and procedures for overriding automated routing.

Another concern is that deferral policies can hide model weaknesses. A system may appear safe because it defers many difficult cases, but this may impose unsustainable workload on clinicians. For that reason, this paper reports review budget, review yield, queue pressure, and delay penalty together. A selective system should not be evaluated only by accepted-case accuracy. It should also be evaluated by what it sends to humans and what burden that creates.

The results support a broader shift in evaluation language. Instead of asking whether a model is accurate enough in the abstract, one can ask what fraction of cases can be automated at a specified harm tolerance and review capacity. This framing is more operational. It allows a model with imperfect accuracy to be evaluated within a defined clinical role. It also makes explicit that the same model may be acceptable under one capacity condition and unacceptable under another.

Future work should replace the simulated components with measured workflow data. Real studies should estimate urgency distributions, radiologist review times, disagreement rates, and downstream delay costs. They should also evaluate open-ended answers, multi-label reports, and cases where human review is not perfectly reliable. The conformal calibration procedure could be extended to handle distribution drift by updating quantiles over time. The queue model could be replaced with discrete-event simulation or hospital-specific scheduling data.

The proposed framework is not limited to emergency radiology. Similar deferral problems occur in pathology, dermatology, ophthalmology, and clinical note summarization [27]. In each setting, automated outputs have unequal harm, and expert review has limited capacity. The exact harm weights and calibration groups would differ, but the core idea remains: uncertainty should be connected to decision cost before it determines whether automation is accepted.

8. Conclusion

This paper presented a simulation-based empirical study of deferral policies for medical vision-language question answering under limited radiologist review capacity. The proposed harm-weighted conformal policy combined conditional nonconformity calibration, urgency-sensitive error costs, and queue-aware review allocation. It was compared with confidence thresholding and entropy-based abstention under matched review budgets across simulated emergency chest imaging cases.

The main result was that harm-aware routing reduced expected clinical penalty without increasing review volume. At a 30% review budget, the proposed method produced lower penalty than both baselines and improved sensitivity for high-urgency positive cases. Its advantage was largest when confidence was miscalibrated across urgency strata and when acute cases were more common. These findings show that selective automation can look different when evaluated through weighted operational loss rather than raw error.

The study also showed that accepted-case accuracy is not enough to judge a deferral system. The proposed method sometimes accepted more low-harm errors than the baselines, yet it reduced total penalty by preventing more high-harm errors. This trade-off is appropriate only if the harm weights and urgency estimates are clinically justified. The result therefore supports transparent decision modeling rather than automatic trust in any routing rule.

The work has limitations. The data are simulated, radiologist review is simplified, and the answer model is represented through probability distributions rather than a deployed generative system. The numerical results should not be treated as clinical performance estimates. They are evidence that the structure of the deferral rule can materially affect workflow-level outcomes under controlled conditions. Real-world validation would require prospective measurement, local calibration, and clinician review of routing priorities [28].

The broader conclusion is that medical vision-language evaluation should include the decision environment in which the model is used. When human review is limited, the question is not only whether the model can answer correctly. It is also which answers should be accepted, which should be reviewed, and how review capacity should be allocated when mistakes have unequal consequences. A risk-calibrated deferral framework provides one way to study that problem with explicit assumptions and

measurable trade-offs.

Acknowledgments

We would like to thank the reviewers of this research for their helpful comments and suggestions.

References

[1] E. Keles and U. Bagci, "The past, current, and future of neonatal intensive care units with artificial intelligence: a systematic review.," *NPJ digital medicine*, vol. 6, pp. 220–220, 11 2023.

[2] B. Sadanandan and V. Behzadan, "Consistent but dangerous: Per-sample safety classification reveals false reliability in medical vision-language models," *arXiv preprint arXiv:2603.20985*, 2026.

[3] L. T. Majnarić, F. Babič, S. O'Sullivan, and A. Holzinger, "Ai and big data in healthcare: Towards a more comprehensive research framework for multimorbidity.," *Journal of clinical medicine*, vol. 10, pp. 766–, 2 2021.

[4] G. Margetis, K. Valakou, S. Ntoa, D. Gavgiotaki, and C. Stephanidis, "Adaptive augmented reality user interfaces for real-time defect visualization and on-the-fly reconfiguration for zero-defect manufacturing.," *Sensors (Basel, Switzerland)*, vol. 25, pp. 2789–2789, 4 2025.

[5] A. Bhardwaj, S. Kishore, and D. K. Pandey, "Artificial intelligence in biological sciences.," *Life (Basel, Switzerland)*, vol. 12, pp. 1430–1430, 9 2022.

[6] A. Cesaro, S. C. Hoffman, P. Das, and C. de la Fuente-Nunez, "Challenges and applications of artificial intelligence in infectious diseases and antimicrobial resistance.," *npj antimicrobials and resistance*, vol. 3, pp. 2–2, 1 2025.

[7] S. Vahdati, E. Mahmoudi, A. Ganjizadeh, C. Chao, and B. J. Erickson, "Decoding large language models for radiology: strategies for fine-tuning and prompt engineering.," *Radiology advances*, vol. 2, pp. umaf024–umaf024, 7 2025.

[8] M. Kılıç, M. Bıyıklı, S. T. A. Özçelik, H. Üzen, and H. Fırat, "Deep learning-assisted detection and classification of thymoma tumors in ct scans.," *Diagnostics (Basel, Switzerland)*, vol. 15, pp. 3191–3191, 12 2025.

[9] Z. Cheng, Y. Wu, Y. Li, L. Cai, and B. Ihnaini, "A comprehensive review of explainable artificial intelligence (xai) in computer vision.," *Sensors (Basel, Switzerland)*, vol. 25, pp. 4166–4166, 7 2025.

[10] A. Whata, K. Dibeco, K. Madzima, and I. Obagbuwa, "Uncertainty quantification in multi-class image classification using chest x-ray images of covid-19 and pneumonia.," *Frontiers in artificial intelligence*, vol. 7, pp. 1410841–1410841, 9 2024.

[11] H. Wang, J. Li, and H. Dong, "A review of vision-based multi-task perception research methods for autonomous vehicles.," *Sensors (Basel, Switzerland)*, vol. 25, pp. 2611–2611, 4 2025.

[12] Y. Li, X. Liu, J. Zhou, F. Li, Y. Wang, and Q. Liu, "Artificial intelligence in traditional chinese medicine: advances in multi-metabolite multi-target interaction modeling.," *Frontiers in pharmacology*, vol. 16, pp. 1541509–1541509, 4 2025.

[13] C. Firestone and B. J. Scholl, "Cognition does not affect perception: Evaluating the evidence for "top-down" effects.," *The Behavioral and brain sciences*, vol. 39, pp. 1–72, 7 2015.

[14] Z. Yao, R. Xi, T. Zhang, Y. Zhao, Y. Tian, and W. Hou, "Construction and enhancement of a rural road instance segmentation dataset based on an improved stylegan2-ada.," *Sensors (Basel, Switzerland)*, vol. 25, pp. 2477–2477, 4 2025.

[15] M. Graña, "Basque conference on cyber physical systems and artificial intelligence," *Zenodo (CERN European Organization for Nuclear Research)*, 5 2022.

[16] S. Brogi and V. Calderone, "Artificial intelligence in translational medicine," *International Journal of Translational Medicine*, vol. 1, pp. 223–285, 11 2021.

[17] M. T. Bianchi, "Integrated welfare systems and disclosure: Approaching emerging issues," *Journal of Modern Accounting and Auditing*, vol. 13, 8 2017.

[18] S. Jamshidiha, A. Rezaee, F. Hajati, M. Golzan, and R. Chiong, "An explainable transformer model for alzheimer's disease detection using retinal imaging.," *Scientific reports*, vol. 15, pp. 26773–26773, 7 2025.

[19] H. Zhu, T. Li, and B. Zhao, "Statistical learning methods for neuroimaging data analysis with applications.," *Annual review of biomedical data science*, vol. 6, pp. 73–104, 4 2023.

[20] I. S. Ahmad, J. Dai, Y. Xie, and X. Liang, "Deep learning models for ct image classification: a comprehensive literature review.," *Quantitative imaging in medicine and surgery*, vol. 15, pp. 962–1011, 12 2024.

[21] P. Becherucci and M. Masoni, "The pediatrician and the digital clinic," *Italian journal of pediatrics*, vol. 45, pp. 3–4, 12 2019.

[22] F. Razemba, L. Jacobs, and D. Franszen, "Convergent validity of the occupational therapy adult perceptual screening test (ot-apst) with two other cognitive-perceptual tools in a south african context," *South African Journal of Occupational Therapy*, vol. 47, pp. 3–10, 8 2017.

[23] M. Aly and N. S. Alotaibi, "A comprehensive deep learning framework for real time emotion detection in online learning using hybrid models.," *Scientific reports*, vol. 15, pp. 42012–42012, 11 2025.

[24] T. E. Mofokeng, "Switching costs, customer satisfaction, and their impact on marketing ethics of medical schemes in south africa: An enlightened marketing perspective," *Cogent Business & Management*, vol. 7, pp. 1811000–, 1 2020.

[25] S. Sun, M. E. Alkahtani, S. Gaisford, A. W. Basit, M. Elbadawi, and M. Orlu, "Virtually possible: Enhancing quality control of 3d-printed medicines with machine vision trained on photorealistic images.," *Pharmaceutics*, vol. 15, pp. 2630–2630, 11 2023.

[26] Y. S. Taspinar, "Machine learning-driven e-nose-based diabetes detection: Sensor selection and feature reduction study.," *Sensors (Basel, Switzerland)*, vol. 25, pp. 6607–6607, 10 2025.

[27] F. H. Awad, M. M. Hamad, and L. Alzubaidi, "Robust classification and detection of big medical data using advanced parallel k-means clustering, yolov4, and logistic regression.," *Life (Basel, Switzerland)*, vol. 13, pp. 691–691, 3 2023.

[28] C. Draheim, R. Pak, A. A. Draheim, and R. W. Engle, "The role of attention control in complex real-world tasks.," *Psychonomic bulletin & review*, vol. 29, pp. 1143–1197, 2 2022.