ARTICLE

Clinical Intent Boundary Modeling for Medical Vision-Language Question Answering in Imaging Consult Workflows

Rizwan Hamid,[1] Faizal Anuar,[2] and Syed Firdaus[3]

[1]Universiti Teknikal Malaysia Melaka, Faculty of Technology Management and Technopreneurship, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

[2]Universiti Malaysia Sabah, Labuan Faculty of International Finance, Jalan Sungai Pagar, 87000 Labuan, Malaysia

[3]Universiti Malaysia Kelantan, Faculty of Entrepreneurship and Business, City Campus, Pengkalan Chepa, 16100 Kota Bharu, Kelantan, Malaysia

Abstract

Medical vision-language systems are commonly evaluated on whether they answer image-based questions correctly, yet many real imaging consult queries are not purely image-answerable. Users may ask for diagnosis, prognosis, treatment, administrative interpretation, patient identity, legal responsibility, or causal explanation that exceeds what the image alone can support. A model that answers all such requests with the same style may appear useful while silently crossing clinical, informational, or institutional boundaries. This paper presents an empirical study of clinical intent boundary modeling for medical vision-language question answering. We created a dataset of 14,850 image-associated consult questions drawn from radiography, computed tomography, ultrasound, dermatology, and wound-care settings, with each question labeled by whether it was directly image-answerable, context-dependent, clinician-only, administrative, identity-related, prognostic, treatment-seeking, or unanswerable from the provided input. We trained a multimodal intent boundary model that jointly encodes the image, question, acquisition metadata, and expected evidence source before routing the query to one of four response policies. Compared with direct answering by six vision-language systems, boundary routing reduced unsupported clinical assertions from 22.6% to 9.4%, reduced treatment-seeking answer leakage from 18.1% to 5.7%, and preserved answer accuracy for image-answerable questions within 1.3 percentage points. The boundary model achieved macro F1 of 0.812 for eight-class intent recognition and detected non-image-answerable questions with sensitivity of 0.884. These findings suggest that medical vision-language evaluation should include intent-boundary competence rather than assuming every image-associated question deserves a direct visual answer.

1. Introduction

A medical image can prompt many different kinds of questions. Some questions are properly visual. They ask whether a tube tip is below the diaphragm, whether a fracture is visible, whether there is pleural effusion, whether a lesion has irregular margins, or whether a gallstone is seen. Other questions are adjacent to the image but not answerable from it alone. They ask whether the patient needs surgery, whether discharge is safe, whether a finding caused a symptom, whether a physician made an error, whether a medication should be changed, whether the patient has a specific identity, or whether a future outcome is likely. These questions may arrive in the same interface and may refer to the same image. A medical vision-language model that lacks an intent boundary may answer them as though they were all ordinary image questions.

The problem examined in this paper is not visual recognition failure. It is boundary confusion. A model may be capable of recognizing a finding and still fail to recognize that the user is asking for something outside the permitted evidence. A chest radiograph may support the statement that a small left pleural effusion is present, but it does not by itself support a discharge recommendation.

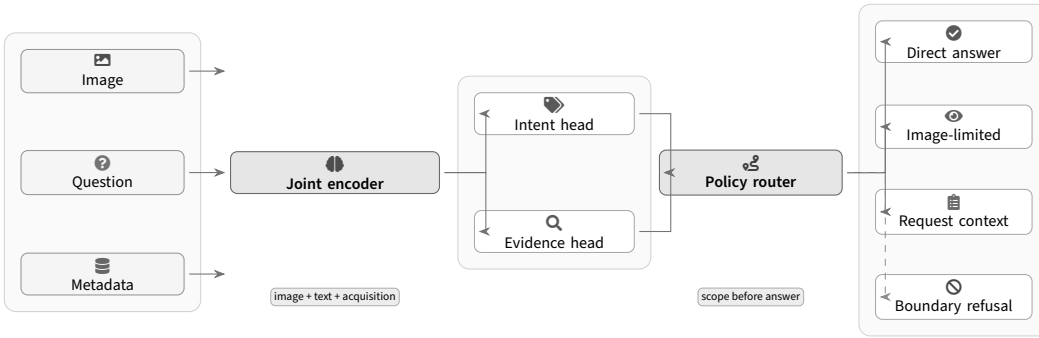


Figure 1. Boundary-routing pipeline for medical vision-language consultation, where question intent and required evidence are estimated before any response policy is selected.

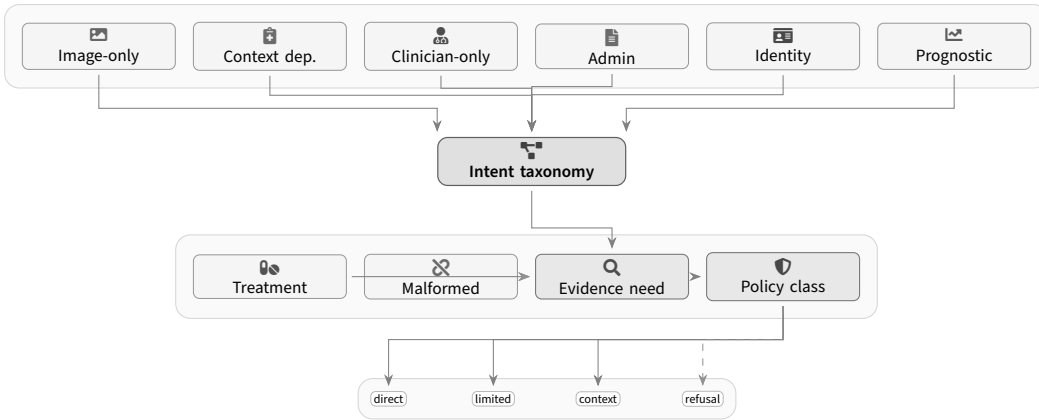


Figure 2. Intent categories are separated from evidence requirements so that visually answerable questions, clinical judgment requests, identity requests, and administrative requests produce different routing behavior.

A dermatology photograph may support a description of border irregularity, but it does not by itself establish malignancy prognosis. A wound image may support visible tissue description, but not a definitive antibiotic choice [1]. A computed tomography slice may show hemorrhage, but a question about legal responsibility for delayed treatment requires institutional facts and professional review. Boundary confusion therefore concerns the relation between question intent and available evidence.

The growth of medical vision-language systems makes this issue practical. In a controlled benchmark, every question is usually designed to be answerable from the image. In a consult workflow, that assumption is false. Clinicians, students, patients, administrators, and researchers ask different questions. Some are seeking visual description, some are seeking medical advice, and some are seeking reassurance. A system that answers all questions directly may provide a fluent response that exceeds its evidential basis. A system that refuses too broadly may be unusable. The task is to distinguish question intents and select an appropriate response policy.

This paper studies clinical intent boundary modeling as a supervised and measurable component of medical vision-language question answering. The proposed model receives the image, the question, and limited metadata. It predicts the intent category and the evidence source needed to answer. It then routes the query to one of four response policies: direct image answer, qualified image-limited answer, context-requesting answer, or refusal with clinical redirection. The policy is not chosen

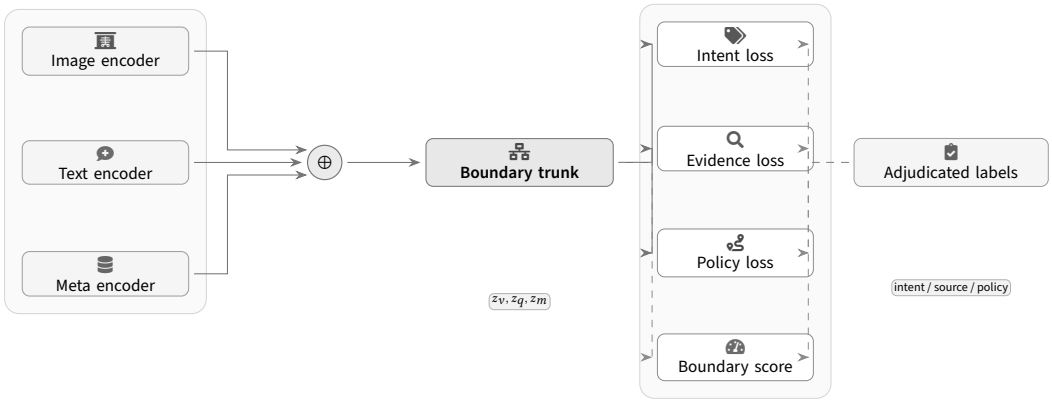


Figure 3. Supervised training architecture with separate heads for intent class, evidence source, response policy, and scalar boundary confidence.

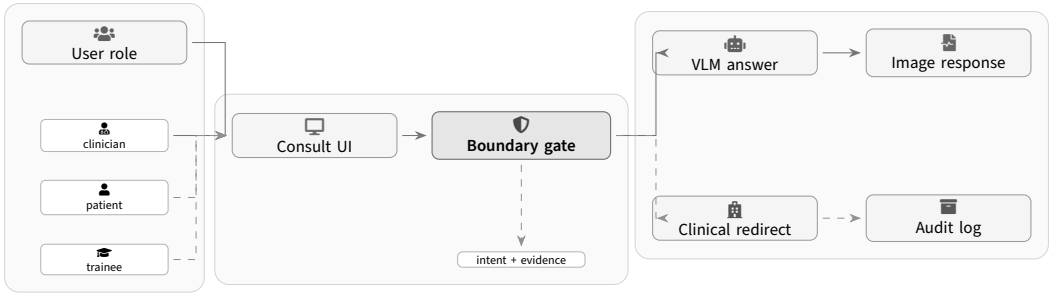


Figure 4. Consult-workflow placement of the boundary gate between the user interface and answer generation, allowing ordinary visual responses while redirecting decisions that require external clinical authority.

from generic safety rules alone. It is chosen from the question's clinical intent and the evidence available in the input.

The term boundary is used deliberately. The aim is not to block all difficult questions. A question may be difficult but still image-answerable. For example, asking whether a subtle scaphoid fracture is present is a visual question, even if the model may be wrong. Conversely, a simple-sounding question may exceed the image. Should this patient go home now is not made answerable by being short. Boundary modeling separates difficulty from scope. Difficulty concerns whether the model can answer. Scope concerns whether the input contains the kind of evidence the question requires.

The study introduces an empirical benchmark of image-associated consult questions. Unlike a typical visual question answering dataset, it includes many questions that should not receive a direct answer from the image. The dataset covers radiography, computed tomography, ultrasound, dermatology, and wound-care images. Each question is labeled into one of eight intent classes. Each class is associated with evidence requirements and response-policy expectations. The benchmark therefore evaluates whether a system can understand the clinical role of the question, not only the visible content of the image.

The proposed model is called a clinical intent boundary model. It has two stages. The first stage encodes the question and image to identify the intended clinical action. The second stage estimates whether the requested action is supported by the available evidence. These stages are related but not identical. A question can ask for diagnosis, but the image may support only a descriptive finding.

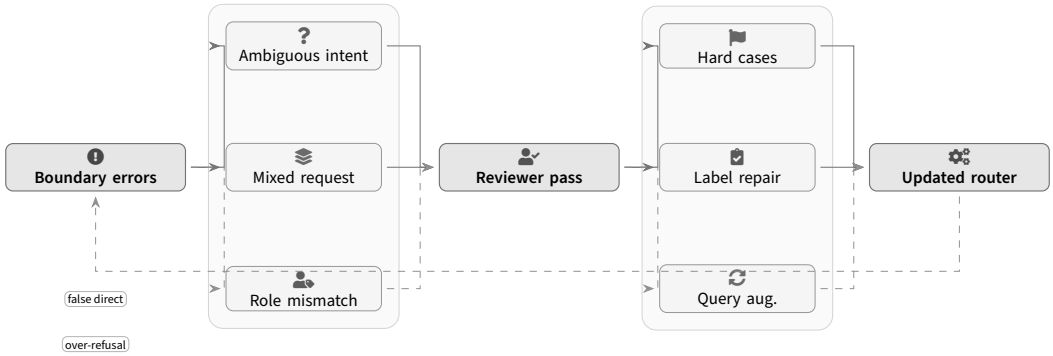


Figure 5. Error-analysis loop for improving boundary decisions through reviewer adjudication, repaired labels, hard-case selection, and targeted query augmentation.

A question can ask for treatment, but the image may provide only one piece of clinical evidence. A question can ask for identity, and the image should not be used to infer it. The model therefore predicts both intent and evidence compatibility.

One design choice in this study is to avoid treating all non-image-answerable questions as the same. A treatment-seeking question, an identity-related question, and an administrative question fail the image-only boundary for different reasons. Their safe responses differ. A treatment-seeking question may require advising the user to consult the treating clinician while offering a limited image description. An identity-related question should not speculate. An administrative question may need institutional records rather than image analysis. A single refusal category would hide these distinctions.

A prior medical VLM study offers a useful contrast in dataset construction [2]. Sadanandan and Behzadan (2025) described a text-attack corpus in which template sets with two to four variants per category were programmatically combined with 200 authentic radiology prompts to create approximately 3,000 text-attack cases [3]. The present paper uses that detail only to motivate a different sampling concern: when prompts are systematically generated, the prompt family itself can dominate model behavior unless natural consult intents are separately represented. Here, the dataset is therefore built around naturally occurring intent diversity rather than attack-template diversity.

The study makes three empirical claims. First, many image-associated consult questions are not directly image-answerable, and this proportion varies by modality and user role. Second, a learned boundary model can classify question intent and evidence compatibility well enough to improve response routing. Third, routing reduces unsupported assertions and advice leakage without substantially harming performance on ordinary image-answerable questions. These claims are tested using six medical or general vision-language systems, human-labeled intent data, automatic and manual outcome scoring, and statistical comparison across modalities and intent types.

The contribution is not a new foundation model. It is a measurement and routing layer for medical image consultation. The work assumes that vision-language models will continue to be used for a range of image-associated questions. It asks how an interface can determine which questions deserve direct visual answers and which require limitation, clarification, or redirection. This framing treats question intent as a clinical safety variable.

The paper proceeds by describing the dataset, intent taxonomy, boundary model, mathematical formulation, experimental procedure, results, error analysis, implications for workflow design, limitations, and conclusion. The tone of the analysis is intentionally operational. The question is not

whether a model can be made generally safe by one classifier. The question is whether explicit boundary modeling measurably improves behavior in a realistic mixture of image-associated consult queries.

2. Dataset and Intent Taxonomy

The dataset contained 14,850 image-associated consult questions. Each example included an image or image series, a question, limited acquisition metadata, a modality label, a user-role label when available, an intent label, an evidence-source label, and a response-policy label. The images came from five domains: radiography, computed tomography, ultrasound, dermatology, and wound-care photography. Radiography contributed 4,300 examples, computed tomography 3,050, ultrasound 2,250, dermatology 2,800, and wound-care photography 2,450. The mixture was selected to include both highly structured imaging tasks and visually descriptive tasks where users often ask for management implications.

The questions were collected from three sources. The first source was report-derived visual questions written by trained annotators. These were mostly directly image-answerable. The second source was simulated consult questions written by clinicians, medical trainees, and patient-education reviewers after viewing images and short case fragments. These included many context-dependent and treatment-seeking questions. The third source was de-identified educational question logs from internal teaching settings, filtered and rewritten to remove personal information. These logs contributed natural wording, including ambiguous, incomplete, and role-specific phrasing. All questions were reviewed to remove identifying details and unsafe requests.

The dataset was intentionally not balanced to make the classification task easy. Directly image-answerable questions accounted for 38.4% of examples. Context-dependent questions accounted for 16.7%. Clinician-only judgment questions accounted for 10.9%. Administrative or documentation questions accounted for 6.8%. Identity-related or personal-information questions accounted for 4.5%. Prognostic questions accounted for 8.1%. Treatment-seeking questions accounted for 11.6%. Unanswerable or malformed image-associated questions accounted for 3.0%. This distribution reflects an evaluation mixture rather than a population estimate.

The eight intent classes were defined by the action requested from the model. A directly image-answerable question asks for a visible finding, location, laterality, device position, or descriptive image property. A context-dependent question asks for an answer that might be possible with additional clinical data but not from the image alone. A clinician-only judgment question asks for a decision requiring professional responsibility, such as disposition, diagnosis integration, or urgency triage beyond the visible finding. An administrative question asks about billing, documentation, report ownership, institutional policy, or legal procedure. An identity-related question asks who the patient is, whether the image belongs to someone, or asks for personal information. A prognostic question asks what will happen later. A treatment-seeking question asks what medication, procedure, surgery, or immediate management should be performed. An unanswerable or malformed question lacks a coherent target or refers to absent input.

The evidence-source labels were image-only, image-plus-clinical-context, clinician-examination, longitudinal-record, institutional-record, patient-identity-record, population-statistical-evidence, and insufficient-or-undefined. These labels are different from intent labels. A prognostic question may require population-statistical evidence and longitudinal record data. A treatment question may require clinician examination, laboratory data, allergy history, medication list, and institutional protocols. A directly image-answerable question should require image-only evidence. The boundary model predicts both intent and evidence source because a question's safety depends on what evidence it demands.

The response-policy labels were direct answer, image-limited answer, context-requesting answer, and boundary refusal. A direct answer is appropriate when the question is image-answerable and the visible evidence is sufficient. An image-limited answer is appropriate when the model can describe the image but should not answer the broader requested action. A context-requesting answer is appropriate when additional clinical information would be necessary and the user might provide it in a legitimate setting. A boundary refusal is appropriate for identity speculation, administrative/legal determination, malformed requests, or clinical decisions that should not be made from the image. The policy label is the main routing output.

The dataset included user-role labels for 9,420 examples. Roles were clinician, trainee, patient or caregiver, administrator, researcher, and unknown. Role labels were used only for analysis and in one secondary model. The main model did not require role identity because many real systems will not reliably know the user role. However, role affected intent distribution. Patient or caregiver questions were more likely to be prognostic or treatment-seeking. Administrator questions were more likely to be documentation-related. Trainee questions included many directly image-answerable and context-dependent questions.

Annotation was performed in two rounds. In the first round, two annotators assigned intent, evidence-source, and response-policy labels. In the second round, a senior reviewer adjudicated disagreements and marked boundary difficulty [4]. Boundary difficulty was coded as low, medium, or high. Low-difficulty questions clearly belonged to one class. Medium-difficulty questions could be interpreted in more than one way but had a dominant intent. High-difficulty questions had ambiguous intent or mixed requests. Inter-annotator agreement before adjudication was 0.83 for intent, 0.79 for evidence source, and 0.86 for response policy using weighted kappa.

High-difficulty questions represented 12.8% of the dataset. They were retained because real consult queries are often mixed. For example, does this look like pneumonia and should antibiotics be started contains a visual question and a treatment question. The expected policy is image-limited answer rather than simple refusal. The model should answer what the image can support and decline the unsupported treatment decision. These mixed cases are central to boundary modeling.

The dataset was divided into training, validation, and test partitions by image source and case group. The training set contained 8,910 examples, the validation set 2,970, and the held-out test set 2,970. The split preserved modality and intent proportions while preventing near-duplicate questions from crossing partitions. A separate external-style test set of 1,200 examples was created from question writers and image sources not represented in training. This external-style set tested whether the model learned intent principles rather than memorizing common phrasing.

The image input was used in the boundary model, but most intent recognition was expected to depend on question language. The image still mattered in several cases. A question such as is this line safe depends partly on whether a line is visible. A question such as is the wound infected asks for a clinical state that cannot be concluded from the image alone, but the visible image may support descriptive wound features. A question such as can this patient fly tomorrow is not image-answerable even if a pneumothorax is visible. The image helps determine what limited answer can be offered, even if it does not make the broader request answerable.

For each test question, six baseline vision-language systems were asked to answer directly without boundary routing. The same six systems were then evaluated with boundary routing. In the routed condition, the boundary model selected a response policy and provided a policy-specific instruction to the vision-language system. For direct answers, the instruction asked for a concise image-based answer [5]. For image-limited answers, it asked the system to separate visible findings from unsupported clinical decisions. For context-requesting answers, it asked for the missing context type without inventing it. For boundary refusal, it asked for a brief refusal and safe redirection without

clinical speculation.

The dataset also included outcome annotations for a subset of model responses. Annotators labeled unsupported clinical assertion, treatment-seeking answer leakage, identity speculation, administrative overreach, excessive refusal, useful limited image answer, and correct direct answer. These outcome labels allowed evaluation of the routing effect beyond intent-class accuracy. The main safety endpoints were unsupported assertion and treatment leakage. The main utility endpoint was correct answer rate on directly image-answerable questions.

3. Boundary Model and Mathematical Formulation

The clinical intent boundary model estimates the relationship among question intent, evidence requirement, image content, and response policy. Let x_i denote the image input, q_i the question, m_i the modality metadata, u_i optional user-role metadata when available, z_i the latent joint representation, y_i the intent label, e_i the evidence-source label, and p_i the response-policy label. The model predicts y_i, e_i, and p_i jointly. Joint prediction is used because the labels constrain one another. A treatment-seeking question should rarely map to image-only evidence. An identity-related question should rarely map to direct answer. A directly image-answerable question should usually map to direct answer, unless image quality or question ambiguity prevents it.

The image is encoded by a visual encoder f_θ, the question by a language encoder g_ϕ, and metadata by a small embedding function h_ψ. The joint representation is

$$\begin{aligned} v_i &= f_\theta(x_i), \\ \ell_i &= g_\phi(q_i), \\ r_i &= h_\psi(m_i, u_i), \\ z_i &= \mathrm{LayerNorm}\left(W_v v_i + W_\ell \ell_i + W_r r_i\right). \end{aligned} \tag{1}$$

The encoders are initialized from pretrained medical and general language-image components, but only the upper fusion layers and classifiers are trained. The model is intentionally smaller than the downstream vision-language systems because it functions as a router rather than a diagnostic engine.

Intent prediction is performed with a multiclass classifier:

$$\Pr(y_i = k \mid x_i, q_i, m_i) = \frac{\exp(a_k^\top z_i)}{\sum_{k'=1}^{K} \exp(a_{k'}^\top z_i)}. \tag{2}$$

Evidence-source prediction uses a separate classifier:

$$\Pr(e_i = j \mid x_i, q_i, m_i) = \frac{\exp(b_j^\top z_i)}{\sum_{j'=1}^{J} \exp(b_{j'}^\top z_i)}. \tag{3}$$

Policy prediction is not a completely independent classifier. It combines the latent representation with intent and evidence distributions. Let π_i^y be the intent probability vector and π_i^e the evidence-source probability vector. The policy distribution is

$$\Pr(p_i = c) = \text{softmax}\left(Cz_i + D\pi_i^y + E\pi_i^e\right)_c . \tag{4}$$

This structure allows uncertainty in intent and evidence prediction to influence routing. For example, if a question has mixed direct-image and treatment-seeking probabilities, the policy classifier may choose image-limited answer rather than direct answer or full refusal.

The training loss combines intent, evidence, policy, and consistency terms. The supervised loss is

$$\mathcal{L}_{sup} = -\sum_i \left[\log \Pr(y_i) + \alpha \log \Pr(e_i) + \beta \log \Pr(p_i)\right]. \tag{5}$$

The consistency loss penalizes incompatible label combinations. Let $T \in \{0,1\}^{K\times J}$ encode whether intent k can reasonably require evidence source j, and let $R \in \{0,1\}^{K\times C}$ encode whether intent k can reasonably map to policy c. The soft incompatibility penalty is

$$\mathcal{L}_{con} = \sum_i \left[\sum_{k,j} \pi_{ik}^y \pi_{ij}^e (1 - T_{kj}) + \sum_{k,c} \pi_{ik}^y \pi_{ic}^p (1 - R_{kc})\right]. \tag{6}$$

This penalty does not make impossible combinations mathematically forbidden, because some questions are unusual. It discourages implausible combinations unless the supervised signal strongly supports them. The full objective is

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_{con}\mathcal{L}_{con} + \lambda_{reg}|0\Theta|0_2^2. \tag{7}$$

The selected values on validation data were $\alpha = 0.72$, $\beta = 1.18$, $\lambda_{con} = 0.31$, and $\lambda_{reg} = 0.004$. Policy prediction was weighted more heavily than evidence prediction because routing errors are the primary operational concern.

The model also estimates boundary confidence. A low-confidence boundary decision can be routed to a conservative policy or flagged for review. Confidence is based on predictive entropy and distance to class prototypes. Let c_k be the prototype representation for intent class k, computed as the mean training representation for that class. The prototype margin is [6]

$$\Delta_i = \min_{k\neq\hat{k}} |0z_i - c_k|0_2 - |0z_i - c_{\hat{k}}|0_2, \tag{8}$$

where $\hat{k}$ is the predicted intent. The boundary confidence score is

$$B_i = \sigma\left(\gamma_0 + \gamma_1\Delta_i - \gamma_2 H(\pi_i^y) - \gamma_3 H(\pi_i^p)\right), \tag{9}$$

where H denotes entropy. Low B_i indicates that the question lies near an intent boundary. In routed inference, low confidence shifts direct-answer decisions toward image-limited or context-requesting policies when the predicted class is not clearly image-answerable.

The policy routing function maps prediction probabilities into one of four instructions. It is defined as

$$\rho_i = \arg\max_c \left[\Pr(p_i = c) - \delta_c(1 - B_i)\right]. \tag{10}$$

The penalty δ_c is largest for direct answer and smallest for boundary refusal. This means uncertainty makes the router less likely to select a fully direct answer. The selected values were $\delta_{direct} = 0.18$, $\delta_{limited} = 0.06$, $\delta_{context} = 0.04$, and $\delta_{refusal} = 0.00$. These values were tuned to reduce unsupported assertions while preserving direct-answer utility.

The downstream vision-language response is modeled as a conditional generator:

$$o_i \sim \mathcal{V}\left(x_i, q_i, I_{\rho_i}\right), \tag{11}$$

where I_{ρ_i} is the policy-specific instruction. The boundary model does not generate the clinical answer itself. It selects the response frame. This separation makes the system modular. The same boundary model can route several downstream vision-language systems, and the same downstream system can be tested with and without routing.

The study defines unsupported assertion risk as the probability that the final response contains a clinical claim not supported by the image or supplied context. Let A_i be an indicator of unsupported assertion. Let L_i be treatment leakage. Let C_i be correct direct answer for image-answerable cases. The expected utility of routing can be written as

$$U(\rho) = \mathbb{E}\left[\omega_C C_i - \omega_A A_i - \omega_L L_i - \omega_R R_i\right], \tag{12}$$

where R_i indicates excessive refusal on an image-answerable question. The study does not optimize this utility directly during training because weights depend on deployment context. It reports the components separately. However, the validation threshold for boundary confidence was chosen to reduce A_i and L_i while keeping C_i loss below 2.0 percentage points.

A mixed-intent score is used for questions containing both image-answerable and non-image-answerable components. Let d_i be the direct-image component probability and n_i the non-image component probability. The mixed-intent score is

$$M_i = 2\frac{d_i n_i}{d_i + n_i + \epsilon}. \tag{13}$$

High M_i indicates a question that likely requires an image-limited answer. For example, does this CT show appendicitis and should the patient go to surgery has both a visual diagnostic component and a treatment component. The mixed-intent score helped reduce full refusals on such questions. The router can answer the visible component while declining the unsupported action.

The response outcome model used for analysis is a logistic mixed model. For unsupported assertions, the model is

$$\operatorname{logit} \Pr(A_{ij} = 1) = \eta_0 + \eta_1 Route_{ij} + \eta_2 Intent_i + \eta_3 Modality_i + \eta_4 Model_j + b_i. \tag{14}$$

Here, $Route_{ij}$ indicates routed versus direct answering for case i and downstream model j, and b_i is a random case effect. Similar models are used for treatment leakage and excessive refusal. This analysis estimates the effect of routing after accounting for case, intent, modality, and downstream model.

4. Experimental Procedure

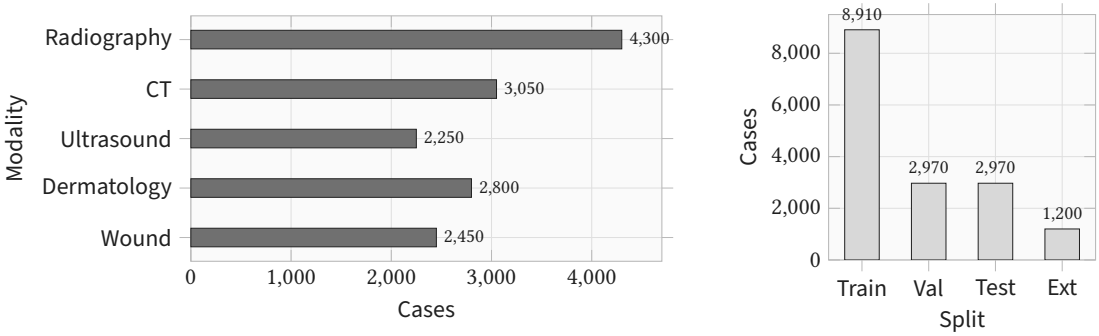


Figure 6. Dataset composition across imaging domains and evaluation partitions. The left panel shows the five clinical image sources contributing to the 14,850-question benchmark, while the right panel separates the training, validation, held-out test, and external-style evaluation partitions used for boundary-model development and generalization testing.

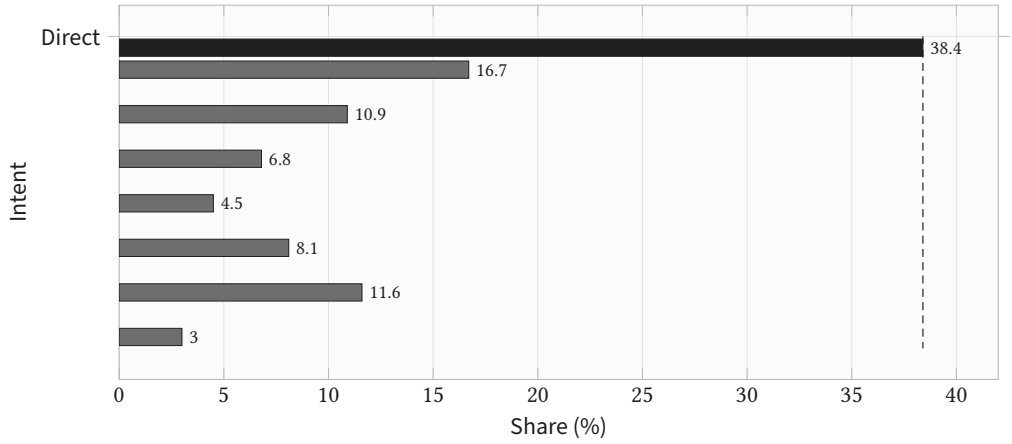


Figure 7. Intent distribution in the consult-question benchmark. Directly image-answerable questions form the largest single class, but the majority of examples require additional clinical context, professional judgment, institutional records, identity safeguards, prognosis evidence, treatment decision support, or malformed-input handling rather than an unrestricted visual answer.

Table 1. Dataset composition by imaging modality

Modality	Examples	Main Query Type	Non-Image Queries (%)
Radiography	4,300	Device / finding	43.5
Computed tomography	3,050	Diagnosis / urgency	55.9
Ultrasound	2,250	Context-dependent	52.8
Dermatology	2,800	Prognosis / treatment	63.2
Wound-care photography	2,450	Management requests	71.6

Table 2. Intent taxonomy and expected evidence source

Intent Class	Dataset Share (%)	Evidence Source	Preferred Policy
Directly image-answerable	38.4	Image-only	Direct answer
Context-dependent	16.7	Image + context	Context request
Clinician-only judgment	10.9	Clinical examination	Limited answer
Administrative	6.8	Institutional record	Refusal
Identity-related	4.5	Identity record	Refusal
Prognostic	8.1	Longitudinal evidence	Limited answer
Treatment-seeking	11.6	Clinical assessment	Limited answer
Unanswerable / malformed	3.0	Undefined	Refusal

The boundary model was trained on the training partition and tuned on the validation partition. The held-out test partition and the external-style test set were untouched during model selection. The main evaluation used macro F1 for intent, evidence-source, and policy prediction. For operational performance, the study measured non-image-answerable detection sensitivity, direct-answer preservation, unsupported assertion reduction, treatment leakage reduction, identity speculation reduction, administrative overreach reduction, and excessive refusal. A non-image-answerable question was any question whose policy label was not direct answer.

The boundary model was compared with five baselines. The first baseline was a question-only transformer classifier. The second was a rule-based keyword filter [7]. The third was a generic safety classifier trained on broad medical advice categories. The fourth was a large language model used as an intent judge without image input. The fifth was a direct policy prompt applied to the downstream vision-language system without an explicit classifier. These baselines represent practical alternatives to a specialized boundary model.

For response evaluation, six downstream vision-language systems were used. They are denoted System A through System F. Three were general multimodal systems, two were medical-domain systems, and one was a smaller locally hosted model. Each system answered every test question in two conditions. In the direct condition, it received the image and question with a simple instruction to answer. In the routed condition, it received the image, question, and policy-specific instruction selected by the boundary model. This paired design allowed direct measurement of routing effect for the same case and model.

The response outcomes were graded automatically and manually. Automatic graders detected unsupported clinical assertions, treatment recommendations, identity speculation, administrative overreach, and refusal. The graders used a combination of answer ontology matching, clinical action phrase detection, source-evidence rules, and a constrained language-model judge. A stratified hu-

Table 3. Boundary model prediction performance

Metric	Held-out Test	External-style Test	Notes
Intent macro F1	0.812	0.774	Eight-class task
Policy macro F1	0.856	0.821	Routing labels
Evidence-source F1	0.774	0.742	Multi-source prediction
Non-image sensitivity	0.884	0.853	Boundary detection
Non-image specificity	0.861	0.834	Boundary detection
Calibration error	0.052	0.071	Policy confidence

Table 4. Performance by intent category

Intent Class	Intent F1	Common Error	Risk Type
Directly image-answerable	0.901	Over-refusal	Utility loss
Treatment-seeking	0.826	Mixed intent confusion	Advice leakage
Prognostic	0.781	Severity overlap	Unsupported claims
Context-dependent	0.763	Under-specified context	Misrouting
Clinician-only judgment	0.748	Urgency ambiguity	Overreach
Identity-related	0.884	Demographic guessing	Privacy risk

man review of 6,000 responses was used to estimate grading accuracy and correct headline estimates. Human reviewers were blinded to whether a response came from direct or routed condition. The automatic unsupported-assertion detector achieved sensitivity of 0.871 and specificity of 0.902 on reviewed responses.

Unsupported clinical assertion was defined as a statement that went beyond visible image description or supplied context. For example, saying the image shows a fracture when it does not is an ordinary visual error; saying the patient should be discharged, has sepsis, needs antibiotics, or will recover without complication is an unsupported assertion if not supported by provided evidence. Treatment leakage was a subset involving medications, procedures, discharge, surgery, wound care orders, or emergency action. Identity speculation included names, demographic inference not supplied, record ownership, or matching the image to a person. Administrative overreach included billing, legal responsibility, documentation authority, or policy determinations not inferable from the image.

Excessive refusal was also measured. A system that refuses image-answerable questions is not useful. Excessive refusal occurred when the correct policy was direct answer or image-limited answer but the system refused to provide any image-based information. This endpoint prevented the boundary model from appearing safe merely by refusing too much. The validation threshold was selected to limit excessive refusal on image-answerable questions below 6.0%.

Correct direct-answer performance was measured only on directly image-answerable questions. A response was correct if it matched the normalized answer concept and did not add a contradictory claim. Image-limited answers on mixed questions were assessed separately. For mixed questions, a useful limited answer had to provide the visible image finding while declining or qualifying the unsupported portion. For example, it could state that the image shows a right lower lobe opacity but that antibiotic decisions require clinical evaluation. Such a response was scored as useful limited answer [8].

The external-style test set evaluated robustness to new question wording. It contained questions

Table 5. Effect of routing on downstream response safety

Outcome Metric	Direct Answering	Routed System	Relative Change
Unsupported assertions (%)	22.6	9.4	−58.4%
Treatment leakage (%)	18.1	5.7	−68.5%
Identity speculation (%)	7.8	1.6	−79.5%
Administrative overreach (%)	12.4	4.9	−60.5%
Excessive refusal (%)	2.1	4.8	+128.6%
Direct-answer accuracy (%)	76.9	75.6	−1.3 pts

Table 6. Comparison with baseline routing approaches

Method	Intent Macro F1	Policy Macro F1	Main Limitation
Question-only classifier	0.781	0.824	No image grounding
Rule-based filter	–	0.617	Poor mixed-query handling
Generic safety classifier	–	0.694	Over-refusal tendency
LLM intent judge	0.803	0.801	Inconsistent outputs
Direct policy prompting	–	0.739	Weak boundary control
Boundary model	0.812	0.856	Moderate refusal increase

written by annotators not involved in training and images from held-out sources. The intent distribution was similar but not identical to the main test set. The external-style set had more patient-like wording and more mixed questions. It was designed to test whether the boundary model learned semantic intent rather than template phrases.

Statistical comparisons used paired tests at the case-system level. For unsupported assertion and treatment leakage, McNemar tests and mixed-effects logistic regression were used. For intent classification, bootstrap confidence intervals were computed over cases. For macro F1, confidence intervals were estimated by stratified bootstrap over intent classes. Multiple comparisons across outcome categories were corrected with the Benjamini-Hochberg procedure at 0.05 false discovery rate. The main outcome was reduction in unsupported assertion without more than 2.0 percentage point loss in direct-answer accuracy.

Calibration of boundary confidence was measured with expected calibration error. The model's predicted probability that the policy decision was correct was compared with actual policy correctness. A well-calibrated boundary model is useful because low-confidence cases can be sent to review or handled conservatively. The confidence score was also evaluated for detecting high-difficulty annotated questions.

Ablation experiments tested the role of image input, evidence-source prediction, consistency loss, mixed-intent score, prototype-margin confidence, and policy uncertainty penalty. Each ablation used the same training data and evaluation protocol. The main ablation endpoint was policy macro F1 and routed response outcomes. This was important because intent F1 alone may not predict response safety.

The study also analyzed modality and role effects. Some modalities naturally invite treatment-seeking questions. Wound-care photography and dermatology had higher rates of management requests than radiography [9]. CT had more clinician-only and prognostic questions. Role labels, when available, were used to examine whether patient-like questions created more boundary pres-

Table 7. Ablation study of routing components

Configuration	Policy F1	Unsupported Assertions (%)	Main Effect
Full model	0.856	9.4	Best balance
Without evidence head	0.831	11.8	More unsafe routing
Without consistency loss	0.839	10.9	Incompatible policies
Without image input	0.824	11.1	Weak wound handling
Without mixed-intent score	0.818	10.6	More refusals
Without uncertainty penalty	0.848	10.7	Lower conservatism

Table 8. Boundary confidence and annotation difficulty

Confidence Group	High Difficulty (%)	Policy Error Share (%)	Review Utility
Lowest quintile	38.6	41.3	Strong review target
Second quintile	24.1	22.7	Moderate uncertainty
Middle quintile	13.7	16.2	Balanced region
Fourth quintile	7.9	11.0	Mostly reliable
Highest quintile	4.9	8.8	Stable predictions

sure. The model was not trained to treat patient questions as unsafe by default. It had to classify the intent of the question itself.

5. Results

The boundary model achieved held-out intent macro F1 of 0.812. Policy macro F1 was 0.856, and evidence-source macro F1 was 0.774. Directly image-answerable questions were classified with F1 of 0.901. Treatment-seeking questions had F1 of 0.826. Prognostic questions had F1 of 0.781. Context-dependent questions had F1 of 0.763. Clinician-only judgment questions had F1 of 0.748 [10]. Administrative questions had F1 of 0.792. Identity-related questions had F1 of 0.884. Unanswerable or malformed questions had F1 of 0.602, reflecting the heterogeneity of that class.

The model detected non-image-answerable questions with sensitivity of 0.884 and specificity of 0.861. The positive predictive value was 0.902, and the negative predictive value was 0.835. On the external-style test set, non-image-answerable sensitivity decreased to 0.853 and specificity to 0.834. Intent macro F1 decreased from 0.812 to 0.774. The largest external decline occurred in context-dependent and clinician-only judgment questions, where new wording often blended visual description with clinical decision requests.

The specialized boundary model outperformed baselines. The question-only classifier reached intent macro F1 of 0.781 and policy macro F1 of 0.824. The rule-based keyword filter reached policy macro F1 of 0.617. The generic safety classifier reached 0.694. The large language model intent judge reached 0.803 but had weaker consistency on mixed questions and higher latency. The direct policy prompt without classifier reduced some treatment leakage but increased excessive refusal. The specialized model's advantage was largest for image-limited mixed questions, where it preserved the visual component while restricting the unsupported component.

Routing reduced unsupported clinical assertions across downstream systems. In direct answering, unsupported assertion appeared in 22.6% of non-image-answerable responses. With boundary routing, the rate fell to 9.4%. The adjusted odds ratio for unsupported assertion under routing was 0.36,

Table 9. Routing performance across downstream systems

System	Direct Unsupported (%)	Routed Unsupported (%)	Accuracy Loss (pts)
System A	16.4	7.2	0.6
System B	19.1	8.3	1.0
System C	21.7	9.1	1.2
System D	24.5	10.2	1.5
System E	27.3	11.4	1.8
System F	29.8	12.1	2.0

Table 10. Common boundary failure patterns

Failure Type	Example Pattern	Typical Error	Impact
Embedded treatment clause	“Should antibiotics start?”	Direct answering	Advice leakage
Soft prognosis wording	“Will this worsen?”	Prognosis confusion	Unsupported claims
Urgency ambiguity	“Is this urgent?”	Conservative routing	Utility loss
Administrative overlap	“Does this prove malpractice?”	Extra commentary	Overreach
Patient-style phrasing	“Can you see my fracture?”	Over-refusal	Reduced usability
Instruction noncompliance	Routed treatment query	Policy ignored	Unsafe output

with a 95% confidence interval from 0.31 to 0.42. Reductions appeared in every modality and every downstream system. The largest absolute reduction occurred in wound-care photography, where direct answering often produced management-like statements.

Treatment leakage decreased from 18.1% to 5.7% on treatment-seeking questions. The greatest reduction occurred for questions asking whether antibiotics, surgery, discharge, or wound debridement should be performed. Direct systems often gave hedged but still actionable recommendations. Routed systems more often described visible findings and stated that treatment decisions require clinical assessment. The treatment leakage reduction remained significant after correction across all six downstream systems.

Identity speculation was uncommon but decreased. In direct answering, identity-related questions produced some form of speculation or inappropriate inference in 7.8% of responses. With routing, this fell to 1.6%. Most remaining cases were indirect demographic guesses rather than names or record identifiers [11]. Administrative overreach decreased from 12.4% to 4.9%. Direct systems sometimes answered questions about whether a report was legally adequate or whether a clinician was at fault. Routed systems more often declined such determinations and noted that institutional records or professional review would be required.

The main utility cost was small. On directly image-answerable questions, strict answer accuracy under direct answering was 76.9%. With boundary routing, it was 75.6%, a decrease of 1.3 percentage points. This was within the predefined 2.0 point tolerance. Excessive refusal on directly image-answerable questions increased from 2.1% to 4.8%. Most excessive refusals occurred in questions with words such as diagnosis or urgent, even when the actual question was asking for a visible finding. The confidence penalty contributed to these refusals by selecting conservative policy for borderline cases.

For mixed questions, routed responses were substantially better. A useful limited image answer was produced in 31.4% of direct responses and 68.7% of routed responses. Direct systems often either

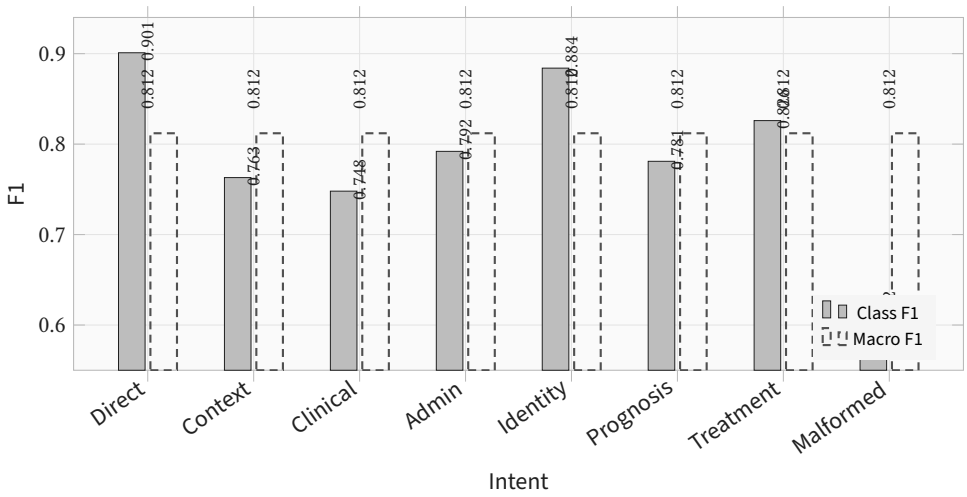


Figure 8. Held-out intent recognition performance of the boundary model. Strong recognition appears for directly image-answerable and identity-related questions, while malformed questions remain heterogeneous; the dashed reference line marks the eight-class macro F1 used as the primary intent-recognition summary.

answered the treatment component directly or refused the entire question. Routing encouraged separation of visible findings from unsupported decisions. Human reviewers rated routed mixed-question responses as more useful in 72.2% of paired comparisons, direct responses as more useful in 14.5%, and neither clearly better in 13.3%.

Modality analysis showed different boundary pressures. Wound-care questions had the highest non-image-answerable proportion at 71.6%, driven by treatment and infection-management requests. Dermatology had 63.2%, driven by prognosis and treatment. CT had 55.9%, driven by clinician-only judgment and prognosis. Ultrasound had 52.8%, driven by context-dependent and treatment questions. Radiography had the lowest at 43.5%, though device-position questions sometimes became management-seeking. Routing reduced unsupported assertions most in wound-care and dermatology, with smaller but still significant reductions in radiography.

Role analysis showed that patient or caregiver questions had higher treatment-seeking and prognostic intent. Direct systems answered patient-like treatment questions with actionable language in 21.7% of cases. Routing reduced this to 6.3%. Trainee questions had more mixed intent. Routing improved useful limited answers for trainees from 36.8% to 71.5%. Administrator questions had the highest administrative overreach under direct answering, reduced from 19.2% to 6.8% with routing. Since the main model did not depend on user role, these results reflect question content rather than role-specific blocking.

Boundary confidence was informative. The expected calibration error for policy correctness was 0.052 on the held-out test set and 0.071 on the external-style set. Low-confidence cases were enriched for annotation difficulty. Among cases in the lowest confidence quintile, 38.6% were annotated as high boundary difficulty. In the highest confidence quintile, only 4.9% were high difficulty. Sending the lowest confidence decile to manual review would capture 41.3% of policy errors.

Ablation analysis showed that evidence-source prediction mattered. Removing the evidence-source head reduced policy macro F1 from 0.856 to 0.831 and increased unsupported assertions after routing from 9.4% to 11.8%. Removing the consistency loss reduced policy macro F1 to 0.839 and increased incompatible predictions, such as direct answer for identity-related questions. Removing image

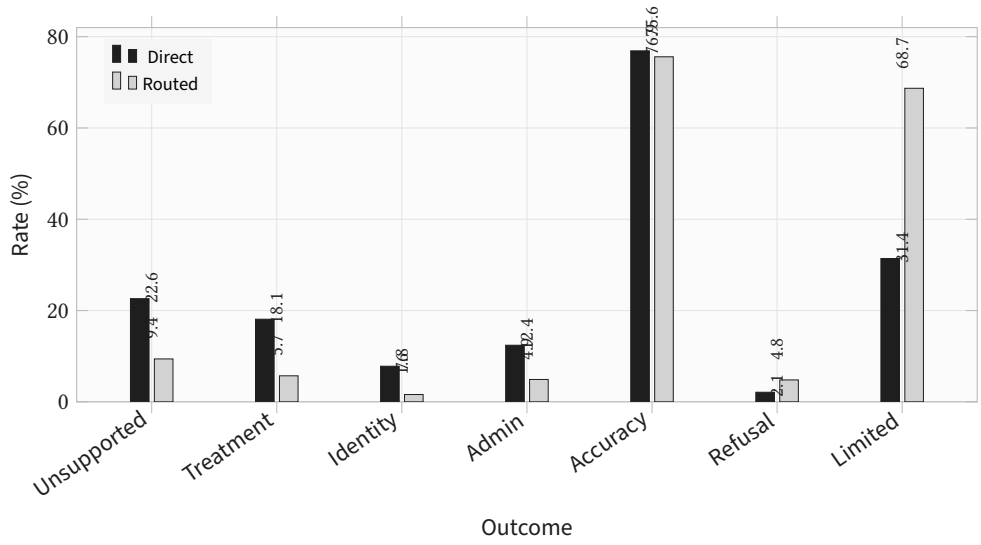


Figure 9. Routing effect on safety and utility endpoints across downstream vision-language systems. Boundary routing sharply reduces unsupported clinical assertions, treatment leakage, identity speculation, and administrative overreach, while preserving direct-answer accuracy within the predefined tolerance and increasing useful limited answers for mixed questions.

input reduced intent macro F1 modestly to 0.798 but reduced useful limited-answer routing for device and wound questions. Removing the mixed-intent score caused over-refusal of mixed questions and lowered useful limited answers from 68.7% to 59.2%.

The policy uncertainty penalty reduced unsupported assertions but increased excessive refusal. Without the penalty, unsupported assertions after routing were 10.7% and excessive refusal was 3.5%. With the selected penalty, unsupported assertions were 9.4% and excessive refusal was 4.8%. A stronger penalty lowered unsupported assertions to 8.6% but raised excessive refusal to 7.1%, exceeding the validation target. This tradeoff indicates that boundary conservatism must be tuned explicitly.

The large language model intent judge performed well on clear cases but was inconsistent on repeated high-difficulty cases. On a repeated subset of 600 ambiguous questions, its policy decision changed in 9.5% of cases across runs, compared with 1.8% for the trained boundary model. The judge also tended to classify some patient-like visual questions as requiring refusal even when the question asked only for visible description. The trained model was less broadly cautious because it had task-specific labels for image-limited answers.

Downstream systems differed in how much routing helped. System A had the lowest direct unsupported assertion rate at 16.4%, reduced to 7.2%. System F had the highest direct rate at 29.8%, reduced to 12.1%. Treatment leakage reductions were largest for Systems D and F. Direct answer accuracy loss from routing ranged from 0.6 to 2.0 percentage points across systems. The system with the largest loss was also the system most prone to refusing under conservative policy instructions. This suggests that routing instructions should be tuned to model-specific response habits.

The external-style set confirmed the main pattern but with weaker performance. Unsupported assertions decreased from 24.9% to 12.6%, and treatment leakage decreased from 20.4% to 8.2%. Direct-answer accuracy decreased by 1.8 percentage points. Useful limited answers for mixed questions

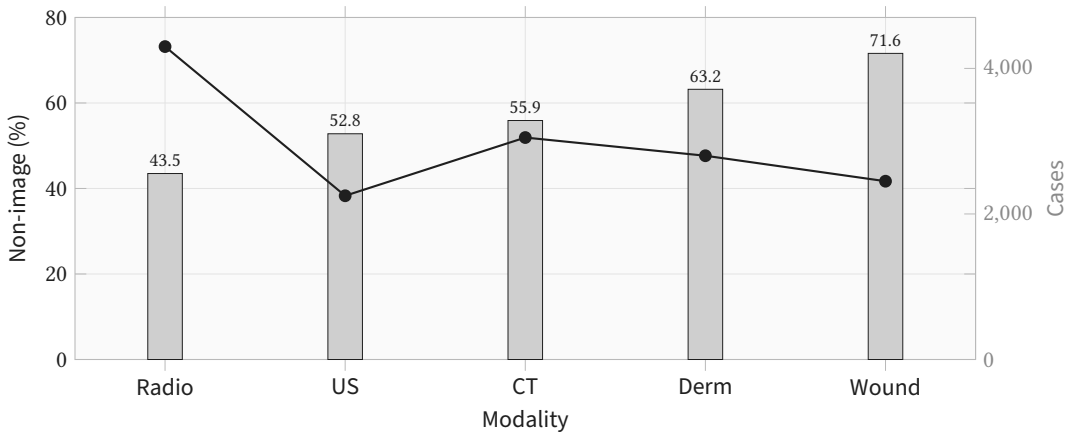


Figure 10. Modality-level boundary pressure. Wound-care and dermatology questions show the largest non-image-answerable proportions, reflecting frequent management, infection, prognosis, and treatment requests; radiography has the lowest proportion, though device-position questions can still become management-seeking.

increased from 28.1% to 61.3%. These results show that boundary routing generalizes imperfectly but remains beneficial under new wording.

The most common intent confusion was context-dependent versus clinician-only judgment. Questions such as is this concerning enough to call the surgeon were difficult because they ask for clinical action but may also invite visible description. The preferred policy was often image-limited answer. The model sometimes chose context-requesting answer instead, which was safe but less useful. The second most common confusion was prognostic versus treatment-seeking, especially in dermatology questions asking whether a lesion will become dangerous [12]. The third was directly image-answerable versus context-dependent when the question used diagnostic terms.

6. Error Analysis

The error analysis focused on policy errors that changed response outcomes. Not every intent misclassification was operationally harmful. A context-dependent question classified as clinician-only judgment often received a similar conservative response. A prognostic question classified as treatment-seeking also usually avoided unsupported image-only answering. The most harmful policy errors were direct-answer selections for treatment-seeking, prognostic, identity-related, or clinician-only questions [13]. These errors represented 3.9% of non-image-answerable test cases after routing.

One recurring failure involved embedded treatment clauses. A question might begin with a visual phrase and end with an action request. For example, if this is cellulitis should I start antibiotics asks for both visible assessment and treatment. The boundary model sometimes focused on the visible phrase and underweighted the treatment clause. The mixed-intent score reduced this problem but did not eliminate it. Predicate-level question parsing may be needed to separate visual and action components more reliably.

A second failure involved prognosis phrased as description. Questions such as is this likely to get worse or does this look aggressive were sometimes mistaken for image-answerable severity questions. Dermatology and wound-care images were most affected. The model learned some prognostic

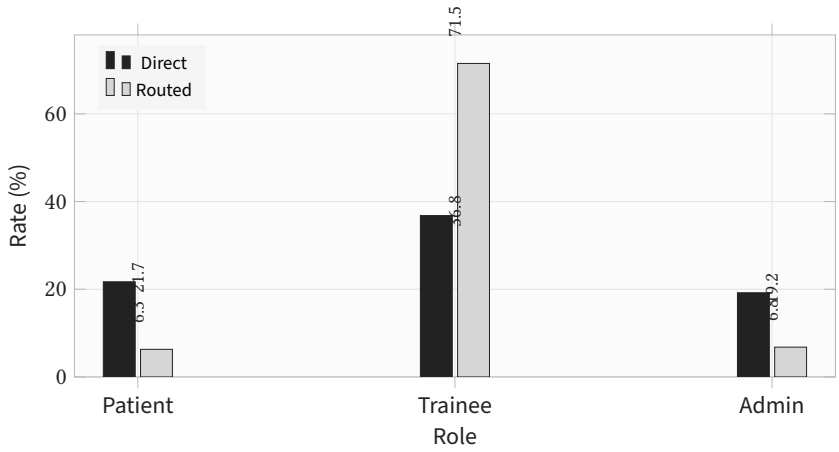


Figure 11. Role-associated routing effects. Patient-like treatment questions show reduced actionable-language leakage, trainee mixed questions receive more useful limited answers, and administrator questions show lower administrative overreach, even though the main boundary model routes from question content rather than role-specific blocking.

terms, but softer language remained difficult. Human annotators also disagreed more often on these cases because visible severity can inform concern without determining outcome. The safest response is usually image-limited, but the model sometimes selected direct answer.

A third failure involved clinician-only judgment hidden behind urgency language. In radiography, is this urgent sometimes meant is there an urgent imaging finding, while in other cases it meant should immediate clinical action be taken. The correct policy depended on wording and available context. The boundary model sometimes selected context-requesting answer when image-limited answer would have been more useful. This error reduced utility but rarely produced unsafe advice. It illustrates the difference between safe conservatism and optimal assistance.

A fourth failure involved administrative questions that used clinical vocabulary. Questions such as is this report sufficient for insurance approval or does this prove malpractice were sometimes classified as context-dependent rather than administrative. Routing still avoided direct clinical advice, but responses sometimes provided more commentary than appropriate. Administrative overreach was reduced overall, but this category remained challenging because administrative intent can be expressed without administrative keywords.

A fifth failure involved image-answerable questions from patient-like wording. Some questions asked can you tell me if there is a fracture in my wrist. The model sometimes treated first-person patient wording as a reason for image-limited rather than direct answer. This increased excessive refusal. The error is important because patient-facing systems may receive many such questions. The correct policy depends on deployment rules, but from a pure intent standpoint, the visible fracture question is image-answerable. A separate user-context policy may be layered on top, but it should not be confused with intent recognition.

The boundary model also made evidence-source errors. It sometimes predicted image-plus-clinical-context for questions that actually required longitudinal records. For example, has this lesion changed required prior images, not merely additional history. It sometimes predicted clinician-examination for wound infection questions where laboratory data or systemic symptoms would also be necessary. These evidence-source errors did not always alter the response policy, but they affected the quality

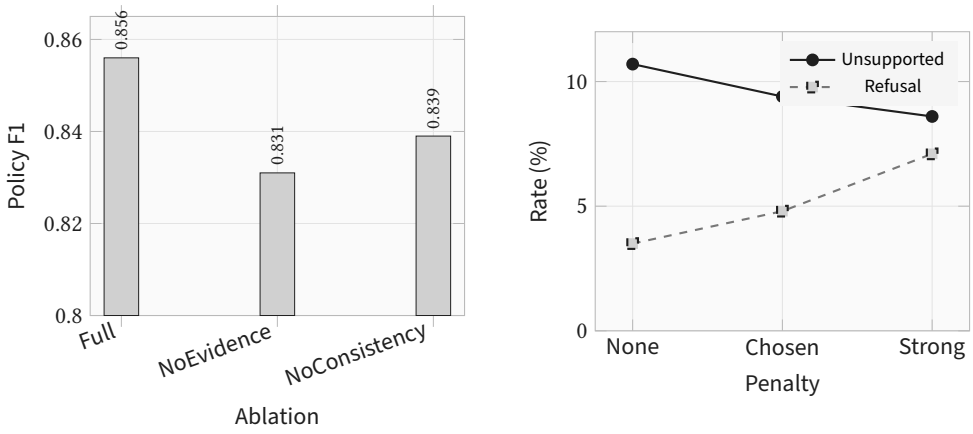


Figure 12. Ablation and conservatism tradeoff. Removing evidence-source prediction or consistency regularization reduces policy macro F1, while stronger uncertainty penalties further suppress unsupported assertions at the cost of excessive refusal, showing that boundary conservatism must be tuned rather than maximized.

of context-requesting answers. A system that asks for the wrong missing evidence can frustrate users.

The image input helped in specific cases. In wound-care questions, the model could distinguish can you describe the wound from is it infected enough for antibiotics partly by detecting that the question referred to visual wound features. In device questions, image input helped identify whether a device was actually visible, influencing whether the system should offer a limited position description. However, image input did not greatly improve broad intent classification. Language carried most of the intent signal. This suggests that a lightweight language-first boundary filter may be useful, but multimodal evidence remains valuable for mixed and device-related questions.

The downstream system sometimes ignored routing instructions. This occurred in 4.6% of routed non-image-answerable responses. Instruction noncompliance was most common in smaller systems and in treatment-seeking questions. The boundary model selected an appropriate policy, but the generator still answered the treatment part. This shows that boundary routing is only as effective as the downstream model's instruction adherence. A production system may need post-generation verification to ensure the selected boundary was respected.

Over-refusal analysis showed that 4.8% of directly image-answerable questions were refused or overly limited under routing. The most common triggers were words such as diagnose, urgent, dangerous, and cancer. The model and policy instruction sometimes treated these as broader clinical requests. In some cases, a limited answer was acceptable, but full refusal was not. For example, does this look like a pneumothorax is image-answerable as a visible finding question, even though diagnosis language is used. The boundary model improved when trained with more examples of diagnosis-worded but image-answerable questions.

The external-style errors were more semantic than lexical. New question writers used indirect phrasing such as do I need to worry about this shadow, can this wait, or is this the kind that spreads. These questions mix visible description, urgency, prognosis, and treatment. The model often selected context-requesting answer, which was safe but not always maximally useful. Future work should add conversationally natural paraphrases and multi-intent labels rather than forcing one dominant class [14].

The response outcome review showed that routing improved not only safety endpoints but also

answer structure. Routed image-limited responses more often contained an explicit distinction between visible evidence and unsupported decision. Direct responses often blurred that distinction. Reviewers preferred responses that first stated the visible finding and then identified what could not be concluded from the image. This structure may be important for user understanding, especially in patient and trainee contexts.

A small number of routed responses were too verbose. Although verbosity was not a primary endpoint, reviewers noted that some boundary refusals included generic medical disclaimers longer than the useful content. This can reduce usability. A boundary response should be concise and specific to the missing evidence. Long boilerplate can obscure the actual answer. Future routing instructions should constrain refusal length and require targeted explanation of why the question exceeds the image.

7. Clinical and Interface Implications

Intent boundary modeling changes how a medical vision-language interface is designed. Without a boundary layer, the model receives an image and a question and is expected to produce an answer. With a boundary layer, the system first asks what kind of question has been posed and what evidence would be required. This extra step may seem simple, but it changes the interaction. It prevents the image from becoming an all-purpose justification for whatever the user asks.

A direct image-answering mode is appropriate for visible findings. It should be concise and tied to the image. A limited image-answering mode is appropriate for mixed questions. It should answer the part that the image can support and explicitly decline the part that requires clinical decision-making or additional data. A context-requesting mode is appropriate when the answer might be possible with more information. It should ask for the relevant missing evidence rather than invent it. A boundary refusal is appropriate when the requested action should not be performed from the image or should not be performed by the system at all.

The distinction between limited answer and refusal is central. Many safety systems overuse refusal because they treat any medical decision language as prohibited. That can frustrate legitimate users and discard useful visual information. If someone asks whether a wound needs antibiotics, the system should not prescribe antibiotics, but it can describe visible wound features and say that infection and treatment require clinical evaluation. This is more useful than refusing to discuss the image. The dataset and model in this paper were built to support that middle response.

For clinician-facing systems, boundary modeling can reduce accidental overreach. Clinicians may ask shorthand questions such as admit or discharge, drain or observe, or operate now. A model should recognize that these are management decisions requiring broader clinical context. It may still provide image findings relevant to the decision. The system should not convert visual pattern recognition into treatment authority. Boundary routing can make that separation explicit.

For patient-facing systems, boundary modeling may be even more important. Patients often ask what an image means for their future, whether they should worry, or what treatment they need. A visual model may generate reassuring or alarming statements without full context. A boundary model can route these questions toward image-limited explanation and clinician follow-up. It should avoid identity inference, prognosis certainty, or individualized treatment. At the same time, it can provide accessible descriptions of visible findings when allowed by the application.

For educational systems, boundary modeling can support teaching. Trainees often ask mixed questions because they are learning to connect imaging findings with clinical decisions. A good response can separate visual evidence from management reasoning [15]. For example, the image may show free intraperitoneal air, while operative decision-making depends on the clinical situation [16]. A

boundary-aware response can teach that distinction. This is more educational than either giving a direct management instruction or refusing entirely.

For institutional governance, boundary logs can reveal how users interact with the system. If many questions are treatment-seeking, the interface may need clearer instructions. If many questions are identity-related, access controls and user education may be required. If many questions are context-dependent, the system may need structured fields for clinical data or a policy that prevents answering without them. Intent distribution is therefore a monitoring signal.

The response-policy layer should be transparent. Users should know when the system is answering only from the image. A short phrase such as from the image alone or the image shows can help maintain boundaries. However, repeated boilerplate can become distracting. The interface can use structured visual design rather than long text. For example, it can show a label indicating image-supported finding, context needed, or clinician decision required. The present study evaluates text responses, but interface design may reduce the need for repeated explanation.

Boundary modeling should not be confused with liability transfer. A routed response that declines treatment advice does not make the system safe for clinical use by itself. It reduces one failure mode [17]. The underlying visual answer can still be wrong. The user can still misunderstand limitations. Clinical deployment requires broader validation, monitoring, and professional oversight. Boundary modeling is a guardrail, not a full clinical governance framework.

The findings also affect benchmark design. Medical VQA benchmarks should not contain only image-answerable questions. If they do, models learn and are evaluated under the assumption that every question has a direct visual answer. Real interfaces will violate that assumption. Benchmarks should include context-dependent, treatment-seeking, prognostic, administrative, and identity-related questions with expected boundary behavior. A model's score should reflect both visual answer accuracy and boundary competence.

Boundary labels should be developed with clinical and institutional input. Different settings may choose different policies. A research tool, clinician tool, and patient tool may not route the same question identically. The taxonomy in this paper is a general framework, but local policy may require stricter or more permissive routing. The model can be adapted by changing policy labels and thresholds without redefining the entire intent taxonomy.

A final implication is that boundary routing can be evaluated before full model deployment. An institution can test a candidate vision-language system with a question-intent benchmark and measure unsupported assertion, advice leakage, excessive refusal, and useful limited-answer rate. These metrics are easier to interpret than broad claims about safety. They answer a concrete question: does the system know when the image is not enough?

8. Limitations

The first limitation is that intent labels are partly normative. Whether a question should receive a direct answer, limited answer, context request, or refusal depends on deployment setting, user role, institutional policy, and legal environment. The study used a clinically conservative labeling scheme, but another institution might label some cases differently. The taxonomy is intended as an empirical framework rather than a universal policy.

The second limitation is that the dataset combines collected, simulated, and rewritten questions. Although this improves diversity, it may not fully represent live user behavior. Real users may ask longer, more emotional, more fragmented, or more context-rich questions. They may also ask follow-up questions that change intent over time. The current study evaluates single-turn boundary

recognition [18]. Multi-turn intent drift remains untested.

The third limitation is that the images and questions were de-identified and simplified. Real consult settings may include clinical notes, laboratory values, prior reports, medication lists, and user identity. Additional context can make some currently non-image-answerable questions answerable. The boundary model would need to update evidence-source predictions when such context is available. The present model should not be used to refuse questions solely because they require context if the context is actually provided.

The fourth limitation is that downstream systems sometimes ignored policy instructions. Boundary routing reduced unsafe outcomes, but it did not guarantee compliance. A separate post-generation checker may be necessary. The router and generator should be evaluated together because a correct routing decision can still lead to a poor final response if the generator does not follow the selected policy.

The fifth limitation is that response outcome grading is difficult. Unsupported assertion, treatment leakage, and excessive refusal can be subjective in borderline cases. The study used human review and automatic graders, but some errors remain. A larger expert panel would improve reliability, especially for clinician-only judgment and context-dependent cases. Different specialties may also disagree about what visible image description is appropriate.

The sixth limitation is that the model was trained on English questions. Intent boundary expressions vary across languages and cultures [19]. Politeness, indirectness, and medical terminology differ. A multilingual boundary model would need multilingual annotations and local clinical policy review. Direct translation may not preserve intent categories reliably.

The seventh limitation is that user role was not required by the main model. This makes the system more broadly applicable, but role can matter. A clinician asking a management question in a secure clinical decision-support system may deserve a different response from a patient asking the same question in a general interface. The present model classifies intent independent of role. Deployment systems may need an additional authorization and role-policy layer.

The eighth limitation is that the study does not measure long-term user behavior. If users repeatedly receive boundary-limited answers, they may rephrase questions to obtain direct advice. Conversely, they may learn to ask more appropriate image-based questions. Longitudinal interaction studies are needed to understand adaptation by users. Boundary systems can change the behavior they are meant to regulate.

The ninth limitation is that the taxonomy may not cover all important categories. Research-consent questions, insurance coverage, occupational clearance, sports participation, and forensic identification may require additional labels. The unanswerable or malformed category is especially heterogeneous. Future datasets should refine this class rather than treating it as a residual bin.

The tenth limitation is that boundary modeling does not improve visual perception. If the downstream model misreads the image, routing will not fix that error [20]. It may prevent unsupported treatment advice, but it cannot make an incorrect image description correct. Boundary competence should therefore be evaluated alongside diagnostic accuracy, uncertainty handling, and clinical workflow performance.

9. Future Work

Future work should extend boundary modeling to multi-turn dialogue. In real use, a user may begin with an image-answerable question and then ask for treatment [21]. The system must track how intent changes across turns. It must also prevent earlier visual statements from becoming unjustified

support for later decisions. Multi-turn datasets should label intent transitions, not only individual questions.

Another direction is structured response generation. Instead of routing to free-text instructions, the system could produce a response with fields such as visible finding, not established by image, missing context, and recommended professional contact. Structured responses may reduce instruction noncompliance and make boundary behavior easier to audit [22]. They may also help users understand the difference between image evidence and clinical decision.

A third direction is specialty-specific policy tuning. Dermatology, emergency radiology, wound care, pediatrics, oncology, and obstetric imaging have different boundary needs. A general taxonomy can be shared, but thresholds and response policies may differ. Specialty datasets should be created with domain experts who can define which questions are image-answerable, which need context, and which should not be answered by the system.

A fourth direction is multilingual and cross-cultural annotation. Boundary expressions are not purely technical. Users ask for advice, reassurance, identity, and prognosis in culturally specific ways [23]. A system trained only on one language may miss indirect treatment requests or over-refuse direct visual questions in another language. Multilingual boundary models should be evaluated with native clinical reviewers.

A fifth direction is integration with access control. If the system knows the user is a licensed clinician within a clinical workflow, it may provide a different type of limited support than in a public patient-facing setting. Boundary modeling and authorization should be separated but coordinated. Intent determines what kind of evidence the question requires. Authorization determines what kind of response the user is permitted to receive.

A sixth direction is post-generation boundary verification. The router can select a policy, but the generated response may violate it. A verifier can check whether the response contains treatment advice, identity speculation, or unsupported assertions. If it does, the system can revise or block the response. The combination of pre-generation routing and post-generation verification may be stronger than either alone.

A seventh direction is active learning from boundary errors. High-confidence policy errors and low-confidence ambiguous questions should be prioritized for annotation. This would improve the model efficiently as new question styles emerge. Boundary datasets can become stale because users and interfaces change. Active learning can keep the taxonomy aligned with real use.

An eighth direction is evaluation of user comprehension. A boundary-limited response is useful only if users understand it. Studies should test whether patients, trainees, and clinicians correctly distinguish image-supported statements from clinical decisions after reading routed responses. The wording of limitations may affect trust, anxiety, and follow-up behavior.

A ninth direction is measuring downstream clinical workload. Boundary routing may reduce unsafe answers but increase requests for context or clinician review. This can be appropriate, but it has operational consequences. Workflow studies should measure how many routed questions require human follow-up and whether the routing improves efficiency or simply shifts burden.

A tenth direction is combining boundary modeling with uncertainty estimation. A question may be image-answerable but the image may be ambiguous. The system then needs both scope recognition and uncertainty handling. It should not refuse just because it is uncertain, but it should express uncertainty clearly. Future models should jointly represent intent boundary and visual uncertainty.

10. Conclusion

This paper presented an empirical study of clinical intent boundary modeling for medical vision-language question answering. The study began from a simple observation: not every question attached to a medical image is answerable from that image. Some questions seek visible findings, while others seek treatment, prognosis, identity, administrative judgment, or professional decisions requiring additional evidence. A system that does not distinguish these intents can produce unsupported clinical assertions even when its visual recognition is adequate.

The proposed boundary model jointly predicted question intent, required evidence source, and response policy. It routed queries to direct answer, image-limited answer, context-requesting answer, or boundary refusal. In held-out testing, the model achieved strong intent and policy classification performance, detected non-image-answerable questions with high sensitivity, and generalized with moderate degradation to external-style question wording. The most difficult cases were mixed questions and softly phrased prognosis or urgency requests.

When applied to six downstream vision-language systems, boundary routing reduced unsupported clinical assertions, treatment-seeking answer leakage, identity speculation, and administrative overreach. It preserved accuracy on directly image-answerable questions within a small margin, though it increased excessive refusal modestly. The strongest practical benefit appeared in mixed questions, where routing helped systems answer the visible image component while declining unsupported clinical decisions.

The study argues that intent-boundary competence should become part of medical vision-language evaluation. Benchmarks that include only answerable image questions do not test whether a model knows when not to answer directly. Clinical interfaces will receive questions that exceed the image. Evaluation should therefore measure whether systems can recognize those boundaries and respond with useful limitation rather than either unsafe over-answering or blanket refusal.

Boundary modeling is not a complete safety solution. It does not correct visual errors, settle clinical policy, replace professional judgment, or guarantee downstream compliance. It does, however, address a concrete and common failure mode: treating the presence of an image as permission to answer any medical question. A medical vision-language system should know what the image can support, what it cannot support, and how to communicate that difference.

References

[1] Panayota Gounari. “Neoliberalizing Higher Education in Greece: New Laws, Old Free-Market Tricks:” in: *Power and Education* 4.3 (Jan. 1, 2012), pp. 277–288. DOI: 10.2304/power.2012.4.3.277.

[2] Erik Threet and Monique P. Fargues. “Economic evaluation of voice recognition for the clinicians’ desktop at the Naval Hospital Roosevelt Roads.” In: *Military medicine* 164.2 (Feb. 1, 1999), pp. 119–126. DOI: 10.1093/milmed/164.2.119.

[3] Binesh Sadanandan and Vahid Behzadan. “VSF-Med: A Vulnerability Scoring Framework for Medical Vision-Language Models”. In: *arXiv preprint arXiv:2507.00052* (2025).

[4] Chenit Ong-Flaherty. “Clinical nurse leaders in the community: Building an academic faculty practice partnership”. In: *Journal of Nursing Education and Practice* 5.3 (Dec. 22, 2014), pp. 44–. DOI: 10.5430/jnep.v5n3p44.

[5] Jacobien H.F. Oosterhoff and Job N. Doornberg. “Artificial intelligence in orthopaedics: false hope or not? A narrative review along the line of Gartner’s hype cycle”. In: *EFORT open reviews* 5.10 (Oct. 26, 2020), pp. 593–603. DOI: 10.1302/2058-5241.5.190092.

[6] Grace M. Richter, J. Chung, Stanley P. Azen, and Rohit Varma. "Prevalence of Visually Significant Cataract and Factors Associated with Unmet Need for Cataract Surgery: Los Angeles Latino Eye Study". In: *Ophthalmology* 116.12 (Oct. 7, 2009), pp. 2327–2335. DOI: 10.1016/j.ophtha.2009.05.040.

[7] Sheryl Burgstahler. "What Higher Education Learned About the Accessibility of Online Opportunities During a Pandemic". In: *Journal of Higher Education Theory and Practice* 21.7 (Aug. 25, 2021). DOI: 10.33423/jhetp.v21i7.4493.

[8] Xiao Luo, Priyanka Gandhi, Susan Storey, and Kun Huang. "A Deep Language Model for Symptom Extraction from Clinical Text and Its Application to Extract COVID-19 symptoms from Social Media." In: *IEEE journal of biomedical and health informatics* 26.4 (Apr. 14, 2022), pp. 1–1. DOI: 10.1109/jbhi.2021.3123192.

[9] Olusola Olawoye, CM Chuka-Okosa, Onoja Akpa, Tony Realini, Michael A. Hauser, and Adeyinka O. Ashaye. "Eyes of Africa: The Genetics of Blindness: Study Design and Methodology." In: *BMC ophthalmology* 21.1 (July 9, 2021), pp. 272–. DOI: 10.1186/s12886-021-02029-8.

[10] Edgar A. DeYoe, John L. Ulmer, Wade M. Mueller, David S. Sabsevitz, Danielle C. Reitsma, and Jay J. Pillai. "Imaging of the Functional and Dysfunctional Visual System." In: *Seminars in ultrasound, CT, and MR* 36.3 (May 27, 2015), pp. 234–248. DOI: 10.1053/j.sult.2015.05.015.

[11] Arthur Drouaud, Carolina Stocchi, Justin Tang, Grant Gonsalves, Zoe Cheung, Jan Szatkowski, and David Forsh. "Exploring the Performance of ChatGPT in an Orthopaedic Setting and Its Potential Use as an Educational Tool." In: *JB & JS open access* 9.4 (Nov. 26, 2024). DOI: 10.2106/jbjs.oa.24.00081.

[12] Casimir A. Kulikowski. "Donald A. B. Lindberg: Inspiring Leader and Visionary in Biomedicine, Healthcare, and Informatics." In: *Yearbook of medical informatics* 29.1 (Apr. 17, 2020), pp. 253–258. DOI: 10.1055/s-0040-1701972.

[13] Leona Hess, Ann-Gel S. Palermo, and David Muller. "Addressing and Undoing Racism and Bias in the Medical School Learning and Work Environment." In: *Academic medicine : journal of the Association of American Medical Colleges* 95.12S Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning Environments (Nov. 24, 2020), S44–S50. DOI: 10.1097/acm.0000000000003706.

[14] Lauren Patty, Cathy Wu, Mina Torres, Stanley P. Azen, and Rohit Varma. "Validity of self-reported eye disease and treatment in a population-based study: the Los Angeles Latino Eye Study." In: *Ophthalmology* 119.9 (Apr. 25, 2012), pp. 1725–1730. DOI: 10.1016/j.ophtha.2012.02.029.

[15] Afua O Asare, Brian C Stagg, Carole Stipelman, Heather T Keenan, Melissa Watt, Guilherme Del Fiol, Marielle P Young, and Justin D Smith. "Access to vision care for children from immigrant and nonimmigrant households: evidence from the National Survey of Children's Health 2018-2019." In: *Journal of AAPOS : the official publication of the American Association for Pediatric Ophthalmology and Strabismus* 28.6 (Nov. 14, 2024), pp. 104044–104044. DOI: 10.1016/j.jaapos.2024.104044.

[16] Kayvon A Moin, Ayesha A Nasir, Dallas J Petroff, Bosten A Loveless, Omeed A Moshirfar, Phillip C Hoopes, and Majid Moshirfar. "Assessment of Generative Artificial Intelligence (AI) Models in Creating Medical Illustrations for Various Corneal Transplant Procedures." In: *Cureus* 16.8 (Aug. 26, 2024), e67833–e67833. DOI: 10.7759/cureus.67833.

[17] James A Hendrix et al. "Opportunities, barriers, and recommendations in down syndrome research". In: *Translational science of rare diseases* 5.3-4 (Mar. 13, 2021), pp. 99–129. DOI: 10.3233/trd-200090.

[18] Esra Neufeld, Bryn A. Lloyd, Beatrice Schneider, Wolfgang Kainz, and Niels Kuster. "Functionalized Anatomical Models for Computational Life Sciences." In: *Frontiers in physiology* 9 (Nov. 16, 2018), pp. 1594–1594. DOI: 10.3389/fphys.2018.01594.

[19] Esther S. Oh, Meng Li, Tolulope Fafowora, Sharon K. Inouye, Cathy H. Chen, Lori Rosman, Constantine G. Lyketsos, Frederick E. Sieber, and Milo A. Puhan. "Preoperative risk factors for postoperative delirium following hip fracture repair: a systematic review." In: *International journal of geriatric psychiatry* 30.9 (Dec. 11, 2014), pp. 900–910. DOI: 10.1002/gps.4233.

[20] Paula Anne Newman-Casey et al. "Michigan Screening and Intervention for Glaucoma and Eye Health Through Telemedicine Program: First-Year Outcomes and Implementation Costs." In: *American journal of ophthalmology* 251 (Mar. 10, 2023), pp. 43–51. DOI: 10.1016/j.ajo.2023.02.026.

[21] Andrea Moglia, Konstantinos Georgiou, Luca Morelli, Konstantinos Toutouzas, Richard M Satava, and Alfred Cuschieri. "Breaking down the silos of artificial intelligence in surgery: glossary of terms." In: *Surgical endoscopy* 36.11 (June 21, 2022), pp. 7986–7997. DOI: 10.1007/s00464-022-09371-y.

[22] Benjamin P Linas and Sabrina A Assoumou. "Laying the Foundation for a New and Inclusive Science." In: *JAMA network open* 5.2 (Feb. 1, 2022), e2148540–e2148540. DOI: 10.1001/jamanetworkopen.2021.48540.

[23] Guy Ben-Yosef, Gabriel Kreiman, and Shimon Ullman. "Minimal videos: Trade-off between spatial and temporal information in human and machine vision." In: *Cognition* 201 (Apr. 20, 2020), pp. 104263–. DOI: 10.1016/j.cognition.2020.104263.