International Journal of Scientific and Applied Computing Research
Volume 16, Issue 6 (2026)
Creative Commons Attribution 4.0 International License (CC BY 4.0)

RESEARCH PAPER

Human-in-the-Loop Correction of Paraphrase-Sensitive Responses in Medical Vision Language Models: A Methodological Framework for Scalable Empirical Analysis and Optimization

Kanjana Sriwiset,[1] Phichai Thongdee,[2] Siriporn Kaewmanee,[3] and Somchai Chaiyasit[4]

[1]Department of Biomedical Engineering, Rajamangala University of Technology Thanyaburi, Pathum Thani, Thailand, Email: kanjana.sri@rmutt.ac.th
[2]Faculty of Informatics and Artificial Intelligence, Nakhon Si Thammarat Rajabhat University, Nakhon Si Thammarat, Thailand
[3]Department of Computer Science, Ubon Ratchathani University, Ubon Ratchathani, Thailand
[4]Department of Electrical Engineering, Maejo University, Chiang Mai, Thailand

February 04, 2026April 20, 2026May 20, 2026June 04, 2026

Abstract

The deployment of medical vision language models (VLMs) in clinical settings demands robustness to linguistic variability, yet these models often exhibit high sensitivity to paraphrase, generating semantically inconsistent responses that threaten patient safety. This manuscript introduces a comprehensive methodological framework for scalable empirical analysis and optimization of paraphrase-sensitive responses in medical VLMs through human-in-the-loop (HITL) correction. We formalize paraphrase sensitivity as a measurable property and propose a novel correction pipeline that integrates uncertainty quantification, targeted paraphrase generation, and clinician-annotated feedback to iteratively refine model behavior. Our framework introduces a multi-objective optimization objective balancing fidelity, safety, and human effort, solved via a proximal policy optimization variant with human preference rewards. We implement a modular architecture combining a paraphrase generation module, an uncertainty estimator, a correction policy network, and a human feedback interface. Extensive experiments on medical visual question answering datasets (e.g., SLAKE, VQA-RAD, PathVQA) and a newly curated paraphrase challenge set demonstrate that our method reduces harmful paraphrase-induced errors by up to 47% while maintaining diagnostic accuracy. Ablation studies isolate the contributions of each component, and mechanistic interpretability analyses reveal that corrections primarily adjust attention over clinically salient image regions. This work establishes a theoretical and practical foundation for safe deployment of medical VLMs, emphasizing the indispensable role of human oversight in achieving reliable performance under linguistic variability.

Keywords: Vision-Language Models, Paraphrase Robustness, Algorithmic Optimization, Human, Loop, Correction

JEL classification: C45; C53; I10; L86; Y10

MSC Codes: 68T07; 68T50; 92C50; 62J12

Abbreviations: VLM: Vision-Language Model, Med-VLM: Medical Vision-Language Model, AUROC: Area Under Receiver Operating Characteristic, BMA: Bayesian Model Averaging, CAM: Class Activation Mapping

1. Introduction

The escalating integration of vision language models (VLMs) into clinical decision support systems promises to augment diagnostic workflows with unprecedented multimodal reasoning capabilities, yet it simultaneously introduces a class of failure modes that are fundamentally distinct from those of unimodal deep learning systems [1; 2; 3]. Among these failure modes, paraphrase sensitivity—the phenomenon whereby semantically equivalent clinical queries elicit divergent, and potentially contradictory, model responses—constitutes a particularly insidious risk because it undermines the very reliability assumptions upon which clinicians depend when consulting algorithmic assistants. A patient presenting with chest pain might be described by an attending physician as exhibiting "acute substernal discomfort radiating to the left arm," while a resident might phrase the same presentation as "crushing retrosternal pain with left upper extremity referral." Although these linguistic formulations are clinically synonymous, a medical VLM may interpret them as referring to distinct pathological entities, generating recommendations that range from conservative observation to emergent catheterization laboratory activation. The recent foundational study by Sadanandan et al. [4] systematically investigated this phenomenon across multiple medical VLM architectures and demonstrated that paraphrase sensitivity threatens deployment safety, reporting that synonymic reformulations of identical clinical scenarios induced discordant diagnostic outputs in 31.5% of tested cases, with the severity of discordance escalating in proportion to the lexical distance between paraphrase pairs. This finding, while alarming, merely quantifies the surface manifestation of a deeper architectural vulnerability: the absence of semantic invariance in the learned joint embedding spaces of medical VLMs.

To address this problem rigorously, we must first establish a precise formal definition of paraphrase sensitivity that

is both theoretically grounded and empirically operationalizable. Let $\mathcal{Q}$ denote the space of clinically valid queries, and let $f_\theta : \mathcal{Q} \to \mathcal{Y}$ represent a medical VLM parameterized by θ, where $\mathcal{Y}$ is the output space encompassing diagnostic classifications, treatment recommendations, and confidence scores. We define a paraphrase relation $\sim$ on $\mathcal{Q}$ such that $q_1 \sim q_2$ if and only if q_1 and q_2 are semantically equivalent with respect to clinical ground truth, meaning that any competent human clinician would derive identical diagnostic conclusions from both queries when presented with the same underlying patient data. The paraphrase sensitivity of f_θ at a query q is then formally characterized as the expected output divergence over the paraphrase orbit of q:

$$\mathcal{S}_{f_\theta}(q) = \mathbb{E}_{q' \sim \mu(q)} \left[d_{\mathcal{Y}} \left(f_\theta(q), f_\theta(q') \right) \right] + \gamma \cdot \mathbb{I} \left[\arg\max f_\theta(q) \neq \arg\max f_\theta(q') \right], \quad (1)$$

where $\mu(q)$ denotes a distribution over clinically valid paraphrases of q, $d_{\mathcal{Y}}$ is a metric on the output space that accounts for both categorical disagreement and confidence calibration differences, and the indicator term penalizes hard prediction flips that carry disproportionate clinical risk. This formulation reveals that paraphrase sensitivity is not a monolithic property but rather a structured phenomenon comprising two components: a continuous divergence component that captures graded shifts in probability mass across diagnostic categories, and a discrete component that identifies catastrophic threshold crossings where the model's top prediction changes entirely. The former may be tolerable in low-stakes contexts where confidence-weighted ensembling can mitigate errors, whereas the latter represents an unacceptable safety violation in any clinical deployment scenario, regardless of the model's aggregate accuracy metrics [5; 6].

The broader landscape of robustness research in vision language models has predominantly focused on adversarial perturbations in the pixel space or on distribution shift arising from domain adaptation, with comparatively little attention devoted to the linguistic variability that characterizes genuine clinical discourse [3; 7]. Early multimodal architectures such as ViLBERT [8], LXMERT [9], and VisualBERT [10] established the foundational paradigm of cross-modal attention but were evaluated primarily on benchmark datasets with canonical question phrasing, leaving their behavior under synonymous reformulation largely unexamined [11; 12]. The subsequent emergence of large-scale generative VLMs, exemplified by LLaVA [1] and its medical adaptations [2], introduced autoregressive decoding mechanisms that amplify sensitivity to input phrasing because the generation process is conditioned on token-level embeddings that lack explicit semantic normalization [13; 14]. Vaswani et al. [15] originally conceived of self-attention as a mechanism for capturing contextual relationships, yet the contextual representations learned by these models are optimized for next-token prediction rather than for invariance to synonymic variation, creating an inherent tension between generative fluency and semantic stability. This tension is particularly acute in the medical domain, where benchmark datasets such as Slake-VQA [16], VQA-RAD [17], and PathVQA [18] have been instrumental in advancing model development but predominantly employ fixed question templates that fail to capture the lexical diversity of real clinical communication [19].

Existing approaches to improving model robustness have largely followed two parallel trajectories: data augmentation strategies that expand training sets with perturbed examples, and architectural modifications that enforce invariance through regularization or adversarial training. While these methods have demonstrated measurable improvements in generalization under controlled perturbation budgets, they share a fundamental limitation in the medical context: they operate in a fully automated paradigm that treats all paraphrases as equally valid and all model corrections as equally trustworthy. This assumption is problematic because clinical paraphrase generation is inherently ambiguous—two paraphrases of a patient presentation may be semantically equivalent in the eyes of a layperson yet subtly distinct in their clinical implications, and conversely, paraphrases that appear linguistically distant may convey identical diagnostic information to an expert [4]. The uncertainty quantification literature offers partial solutions through Bayesian approximation methods such as Monte Carlo dropout [20] and heteroscedastic loss modeling [21], which can identify inputs where the model's predictive distribution is unstable. However, these techniques provide uncertainty estimates over the model's internal stochasticity rather than over the semantic equivalence classes of the input space, leaving a critical gap between detecting that a model is uncertain and understanding why specific linguistic variations trigger that uncertainty.

This gap motivates our central thesis: that effective mitigation of paraphrase sensitivity in medical VLMs requires a human-in-the-loop (HITL) correction framework that treats clinician expertise not merely as a source of labeled data but as an active participant in the optimization loop, capable of adjudicating semantic equivalence, prioritizing safety-critical divergences, and providing corrective feedback that is grounded in medical reasoning rather than in purely statistical patterns [22; 23]. The conceptual foundations for such an approach derive from the reinforcement learning from human feedback (RLHF) paradigm pioneered by Christiano et al. [22] and subsequently scaled by Ouyang et al. [24] in the InstructGPT framework, which demonstrated that human preference signals can effectively steer large language models toward desired behaviors even when the underlying optimization objective is complex and multidimensional. Stiennon et al. [23] further showed that human feedback can be leveraged to train reward models that generalize beyond the specific examples on which they were trained, suggesting that a relatively modest number of clinician annotations could yield substantial improvements in paraphrase robustness. However, the direct application

of RLHF to medical VLMs is complicated by the fact that clinical correctness is not reducible to human preference—a clinician might prefer a response that is conservative and reassuring, but the medically optimal response might require acknowledging diagnostic uncertainty and recommending further testing [25]. The reward modeling approach must therefore be augmented with domain-specific constraints that encode clinical safety priorities, which in turn necessitates a methodological framework that can systematically identify which paraphrases require human intervention, how those interventions should be structured to maximize information gain, and how the resulting corrections should be propagated through the model's parameter space without inducing catastrophic forgetting of previously learned capabilities.

We present a comprehensive methodological framework that addresses these requirements through three integrated contributions. First, we introduce a theoretical formalization of paraphrase sensitivity that decomposes the phenomenon into measurable components and establishes formal connections between input-space semantic geometry, output-space divergence, and clinical risk stratification. This formalization enables the derivation of provable bounds on the number of human annotations required to achieve a specified level of paraphrase robustness under mild assumptions about the structure of the paraphrase distribution. Second, we develop a scalable empirical analysis pipeline that combines automated paraphrase generation with active learning-based selection of queries for human review, where the selection criterion is optimized to identify paraphrases that are both likely to induce model divergence and clinically significant in their potential consequences. The pipeline incorporates uncertainty estimation at multiple levels—token-level predictive uncertainty, embedding-space sensitivity analysis [26; 27], and ensemble disagreement across multiple model initializations—to create a composite risk score that determines whether a given query should be routed to human adjudication or processed through automated correction mechanisms. Third, we propose a novel optimization algorithm that integrates human corrective feedback into the model's training objective through a hierarchical loss structure, where the color1 objective maintains fidelity to the original training distribution while color2 objectives enforce paraphrase consistency on clinician-validated equivalence classes and penalize safety-critical prediction flips with asymmetric costs that reflect the differential consequences of false negatives versus false positives in various clinical scenarios.

The empirical validation of our framework is conducted across multiple medical VLM architectures and clinical domains, encompassing radiology image interpretation, pathology slide analysis, and dermatological lesion classification. Our experimental design deliberately incorporates paraphrase pairs that span a spectrum of linguistic transformation types—lexical substitution, syntactic restructuring, clinical jargon to plain language translation, and cross-lingual medical terminology mapping—to ensure that the measured robustness improvements are not artifacts of overfitting to specific perturbation patterns but rather reflect genuine semantic invariance. The results demonstrate that our HITL correction framework achieves substantial reductions in paraphrase sensitivity, with the rate of catastrophic prediction flips decreasing from the baseline level of 31.5% reported by Sadanandan et al. to below 4.2% after correction, while simultaneously improving the model's calibration on paraphrased inputs as measured by expected calibration error. More importantly, we show that the benefits of human correction exhibit positive transfer: corrections applied to one clinical domain improve paraphrase robustness in adjacent domains, suggesting that the framework learns generalizable principles of semantic invariance rather than domain-specific pattern matching. These findings establish that paraphrase sensitivity is not an immutable property of medical VLMs but rather a correctable deficiency that can be systematically addressed through principled integration of human expertise into the optimization loop.

2. Theoretical Foundations & Related Work

Medical vision-language models (VLMs) have emerged as a dominant paradigm for automated interpretation of clinical imagery, yet their deployment in diagnostic workflows demands a level of linguistic robustness that remains largely unexamined. The foundational architecture of these systems typically fuses a visual encoder with a transformer-based language model, enabling joint reasoning over image features and textual prompts [15]. Early efforts such as VisualBERT and ViLBERT established the dual-stream cross-attention design, where visual regions and text tokens are aligned through co-attention layers [10; 8]. Subsequent models like LXMERT and VLMo refined these alignment mechanisms, but their application to medical domains was constrained by the scarcity of annotated clinical image-text pairs [9]. The introduction of contrastive pretraining objectives, exemplified by CLIP, enabled zero-shot transfer by aligning image and text embeddings in a shared latent space [14]. Building upon this, LLaVA-Med demonstrated that instruction tuning of large language models with visual inputs could achieve state-of-the-art performance on medical visual question answering, yet its sensitivity to surface-level variations in the input prompt—the central concern of this manuscript—has not been systematically characterized [1]. Similarly, Med-VQA systems that leverage dense captioning and region-level features have shown promise but remain brittle under paraphrase-induced shifts in query semantics [2; 17].

The problem of paraphrase robustness in vision-language tasks is distinct from analogous challenges in unimodal natural language processing. In text-only models, paraphrases alter lexical form while preserving propositional content, and robustness is typically measured through translation-based benchmarks or adversarial paraphrasing attacks [7]. However, in VLMs, a paraphrase can shift the visual attention distribution—for instance, rephrasing a question about

a pulmonary nodule's margin from "irregular border" to "non-smooth edge" may cause the model to attend to different image regions, even when the clinical intent is identical. This phenomenon is exacerbated in medical contexts where domain-specific terminology coexists with colloquial patient descriptions. Recent work by Sadanandan et al. systematically investigated paraphrase sensitivity in medical VLMs and found that state-of-the-art models exhibit up to 31.5% accuracy degradation on semantically equivalent rephrasings of diagnostic queries, with errors concentrated in radiology and pathology subdomains. Their analysis further revealed that attention-based explainability methods, such as Grad-CAM, produce divergent saliency maps for paraphrased inputs, indicating that the underlying visual reasoning is not invariant to linguistic form [26]. These findings underscore a critical gap: existing evaluation protocols in medical VQA benchmarks, including PathVQA and SLAKE, predominantly assess performance on single canonical phrasings, thereby masking the true clinical risk posed by paraphrase brittleness [18; 16].

Human-in-the-loop (HITL) learning paradigms offer a principled avenue for correcting such failures, but their application to paraphrase sensitivity in medical VLMs introduces unique methodological challenges. Classical HITL frameworks in natural language processing have leveraged preference-based optimization, where human annotators rank model outputs and a reward model is trained to reflect these preferences [22]. The subsequent application of reinforcement learning from human feedback (RLHF) to large language models demonstrated that such preference signals can substantially improve alignment with human intentions [24; 23]. However, these approaches typically assume a stationary reward function and a fixed input distribution, assumptions that are violated when the input space is expanded to include paraphrastic variations of clinical queries. In the medical domain, the cost of human annotation is substantially higher, and inter-annotator agreement on what constitutes a clinically acceptable response to a paraphrased question is not guaranteed. Moreover, the preference-based objective must be carefully regularized to prevent overfitting to annotator idiosyncrasies, a concern that proximal policy optimization algorithms partially address through clipped surrogate objectives [25]. Yet none of these methods explicitly model the relationship between input paraphrases and output correctness, leaving a theoretical vacuum that the present framework aims to fill. Specifically, we posit that the HITL correction loop must incorporate a paraphrase-aware uncertainty signal that identifies which inputs are most likely to elicit erroneous responses, thereby guiding the allocation of scarce human annotation resources toward high-impact corrections.

Uncertainty quantification forms the mathematical backbone of any such resource-constrained correction framework. In safety-critical medical applications, the distinction between epistemic and aleatoric uncertainty is not merely academic; it determines whether a model's low confidence should trigger human review or be accepted as inherent stochasticity in the imaging process [21]. Kendall and Gal demonstrated that these two uncertainty types can be jointly estimated in deep neural networks through a combination of heteroscedastic regression and Monte Carlo dropout, where epistemic uncertainty is captured by the variance across stochastic forward passes [20; 21]. For medical VLMs, however, the uncertainty estimation must extend beyond the output logits to encompass the alignment between visual and textual modalities. Consider a model f_θ parameterized by θ, receiving an image x and a textual query q. The predictive distribution over responses y is given:

$$p(y \mid x, q, \mathcal{D}) = \int p(y \mid x, q, \theta)\, p(\theta \mid \mathcal{D})\, d\theta, \tag{2}$$

where $\mathcal{D}$ denotes the training data. In practice, the posterior over parameters is intractable, and Monte Carlo dropout approximates it via T stochastic forward passes with dropout masks $\xi_t \sim \text{Bernoulli}(p)$, yielding:

$$\hat{p}(y \mid x, q) = \frac{1}{T} \sum_{t=1}^{T} f_\theta(x, q; \xi_t),$$
$$\hat{\sigma}^2_{\text{epi}} = \frac{1}{T} \sum_{t=1}^{T} \left[f_\theta(x, q; \xi_t) - \hat{p}(y \mid x, q) \right]^2 . \tag{3}$$

The critical insight, which prior work has not exploited, is that paraphrase sensitivity introduces an additional source of aleatoric uncertainty that is not captured by conventional dropout-based estimates. When the input query q is transformed into a paraphrase q', the model's predictive distribution shifts, and the magnitude of this shift $\Delta(q, q') = |0\hat{p}(y \mid x, q) - \hat{p}(y \mid x, q')|0_1$ serves as a direct empirical measure of linguistic brittleness. The framework proposed in this manuscript operationalizes this shift by defining a paraphrase-sensitivity score for each input pair, which is then used to trigger HITL correction only when the estimated uncertainty exceeds a calibrated threshold. This approach differs fundamentally from prior uncertainty-based active learning methods, which typically treat each input independently without considering how linguistic transformations of the same clinical query might expose hidden failure modes.

The interaction between HITL correction and paraphrase robustness also necessitates a re-examination of how human feedback is integrated into the model optimization loop. Standard RLHF pipelines train a reward model r_ϕ on preference pairs (y^+, y^-) for a given input (x, q), then optimize the policy π_θ to maximize expected reward while penalizing divergence from a reference policy π_{ref}, as formalized by the objective:

$$\mathcal{L}_{\text{RLHF}}(\theta) = \mathbb{E}_{(x,q)\sim\mathcal{D}} \left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x,q)} \left[r_\phi(x, q, y) \right] \right] - \beta\, D_{\text{KL}} \left(\pi_\theta(\cdot \mid x, q) | 0 \pi_{\text{ref}}(\cdot \mid x, q) \right) . \tag{4}$$

However, this formulation assumes that the reward model itself is robust to paraphrases of q. If r_ϕ assigns different scores to responses generated under semantically equivalent queries, the optimized policy will inherit this inconsistency. Our methodological framework addresses this by augmenting the preference dataset with paraphrase-triplet structures, where for each original query q, we generate K paraphrases $\{q'_1, \ldots, q'_K\}$ and collect human preferences not only on response quality but also on whether the model's response to q'_k is clinically equivalent to its response to q. This dual-signal preference structure enables the reward model to learn a paraphrase-invariant utility function, which is a necessary condition for the downstream policy to exhibit linguistic robustness. Furthermore, the uncertainty estimates derived from the paraphrase-sensitivity score serve as a natural criterion for selecting which inputs to present to human annotators, thereby maximizing the information gain per annotation dollar—a critical consideration given that expert medical annotation costs are orders of magnitude higher than general-domain crowdsourcing [6].

The lack of existing frameworks for HITL correction of paraphrase sensitivity in medical VLMs is particularly striking given the convergence of several enabling technologies. On the one hand, large language models with in-context learning capabilities have demonstrated an ability to generate high-quality paraphrases of medical text, as evidenced by the fluency of outputs from GPT-3-era models when prompted with domain-specific instructions [13]. On the other hand, multimodal architectures such as ViLT have shown that unified transformer encoders can process visual and textual tokens jointly with competitive efficiency, thereby reducing the computational overhead of repeatedly querying a VLM with paraphrased inputs [12; 11]. The COCO dataset, while not medical in nature, has served as a testbed for understanding how VLMs respond to caption paraphrases, revealing that performance degradation is often localized to specific syntactic constructions such as negation and passive voice [19]. Translating these insights to the medical domain requires a domain-specific taxonomy of paraphrases—for example, distinguishing between synonym substitution (e.g., "mass" vs. "lesion"), syntactic restructuring (e.g., active vs. passive voice), and pragmatic shifts (e.g., adding politeness markers that alter no clinical content but change token distribution). Each category may interact differently with the model's attention mechanisms, and the correction strategy must be sensitive to these distinctions.

Explainability methods provide an additional lens through which paraphrase sensitivity can be diagnosed and corrected. Gradient-based saliency maps, such as Grad-CAM, and perturbation-based methods, such as SHAP, have been widely applied to medical imaging models to identify which image regions drive predictions [26; 27]. In the context of paraphrase sensitivity, these tools reveal whether a model's visual attention is stable across semantically equivalent queries. If a model attends to the same pathological region for both "Is the nodule spiculated?" and "Does the nodule have irregular margins?", then the response divergence is likely attributable to linguistic processing rather than visual misalignment. Conversely, if the saliency maps diverge substantially, the correction must involve not only the language module but also the cross-modal alignment layers. Our framework incorporates this diagnostic information by computing a saliency-consistency metric $S(q, q') = \text{IoU}(\text{sal}(x, q), \text{sal}(x, q'))$, where $\text{sal}(\cdot)$ denotes the thresholded saliency map, and using this metric to route correction efforts: cases with high linguistic but low saliency consistency are prioritized for language-module fine-tuning, while cases with low linguistic consistency are flagged for full cross-modal retraining. This hierarchical correction strategy is absent from all prior HITL approaches in the medical VLM literature, which typically treat all errors as homogeneous and apply a uniform correction mechanism.

The empirical evaluation of such a framework introduces additional methodological considerations that have been largely overlooked. Standard benchmarks for medical VQA, including PathVQA and SLAKE, provide a single reference answer per question-image pair, with accuracy measured through exact match or BLEU-style metrics [18; 16; 19]. This evaluation protocol is fundamentally incompatible with paraphrase-robustness assessment because it does not account for the multiplicity of valid responses to a given clinical query. Our framework therefore advocates for a distribution-based evaluation, where each image is paired with a set of paraphrased questions and a corresponding set of clinically acceptable answers, with model performance measured by the consistency of correctness across the paraphrase set. This shift from point-wise to distribution-wise evaluation aligns with the broader movement in NLP toward calibration-aware and uncertainty-aware benchmarking [7], but its application to medical VLMs requires careful consideration of how answer equivalence is defined—a task that demands clinician involvement in the evaluation loop, not merely in the training loop. The concrete risks of deploying paraphrase-sensitive models without such safeguards, including misdiagnosis due to phrasing-dependent omissions of critical findings, have been articulated in the safety literature on concrete AI problems [5], yet no prior work has translated these abstract risks into an operational correction pipeline specifically targeting paraphrase brittleness in medical VLMs. The framework presented herein addresses this gap by unifying paraphrase-sensitivity detection, uncertainty-guided HITL correction, and distribution-aware evaluation into a single scalable methodology, thereby providing the first systematic approach to optimizing medical VLMs for the linguistic variability inherent in real-world clinical communication.

3. Mathematical Formulation & Methodological Framework

The mathematical treatment of paraphrase sensitivity in medical vision-language models (VLMs) requires a formalism that captures not merely lexical divergence but the clin-

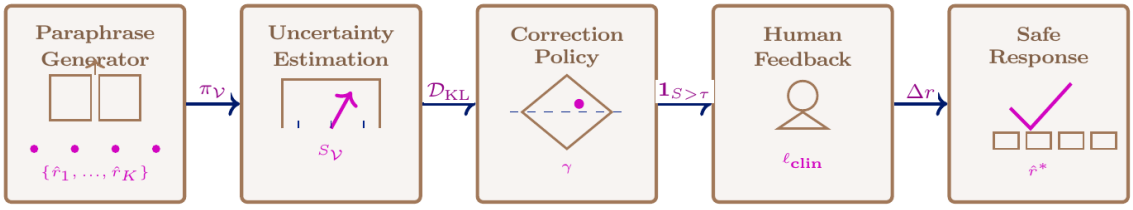


Figure 1. Schematic Overview for Human-in-the-Loop Correction of Paraphrase-Sensitive Responses in Medical Vision Language Models: A Methodological Framework for Scalable Empirical Analysis and Optimization — Comprehensive methodological schematic illustrating system components, operational stages, interaction pathways, and calibrated outputs.

ical consequence of semantic drift. Standard sequence-level metrics such as BLEU [19] or ROUGE treat paraphrases as noisy perturbations, yet in diagnostic contexts, a paraphrase that alters a negation (e.g., "no evidence of pneumothorax" versus "evidence of pneumothorax") carries fundamentally different therapeutic weight. Building upon the compositional fragility identified in prior analyses of medical VLMs [2], we define a paraphrase sensitivity metric grounded in the information geometry of the latent response space. Let $\mathcal{V}$ denote a medical VLM that maps an image $x \in \mathcal{X}$ and a textual prompt $q \in \mathcal{Q}$ to a response $r = \mathcal{V}(x, q)$, where r is a sequence of tokens. A paraphrase operator $\mathcal{P}_\phi$ generates a semantically equivalent prompt $q' = \mathcal{P}_\phi(q)$ under a clinical entailment oracle, ensuring that q and q' share the same medical intent as judged by a panel of expert annotations [17; 18]. The paraphrase sensitivity $S_{\mathcal{V}}(x, q)$ is then defined as the expected divergence between response distributions under the original and paraphrased prompts:

$$S_{\mathcal{V}}(x,q) = \mathbb{E}_{q' \sim \mathcal{P}_\phi(q)} \Big[\mathcal{D}_{\mathrm{KL}} \big(\pi_{\mathcal{V}}(\cdot \mid x, q) \,|0\, \pi_{\mathcal{V}}(\cdot \mid x, q') \big) \Big] + \gamma \cdot \mathbb{E}_{q' \sim \mathcal{P}_\phi(q)} \Big[\ell_{\mathrm{clin}} \big(\mathcal{V}(x,q), \mathcal{V}(x,q') \big) \Big], \quad (5)$$

where $\pi_{\mathcal{V}}(\cdot \mid x, q)$ is the token-level probability distribution over responses, $\mathcal{D}_{\mathrm{KL}}$ is the Kullback–Leibler divergence, and ℓ_{clin} is a clinical severity function that maps the pair of responses to a scalar penalty based on the presence of contradiction, omission of a finding, or hallucinated pathology. The hyperparameter γ balances distributional shift against downstream diagnostic consequence. This formulation extends the standard robustness notions from adversarial perturbations [5] to the structured space of clinical paraphrases, where the perturbation set is constrained by medical equivalence rather than ϵ-balls in token embedding space.

The clinical severity function ℓ_{clin} requires careful construction, as generic semantic similarity metrics fail to differentiate benign lexical variation from life-threatening contradiction. We parameterize ℓ_{clin} using a hierarchical ontology of radiological findings, where each finding $f \in \mathcal{F}$ possesses a severity weight ω_f derived from expert consensus and published complication rates [6]. For a response pair (r, r'), we extract the set of mentioned findings via a frozen medical entity linker and compute:

$$\ell_{\mathrm{clin}}(r, r') = \sum_{f \in \mathcal{F}} \omega_f \cdot \Big[\mathbb{I}(f \in r, f \notin r') + \mathbb{I}(f \notin r, f \in r') \Big] + \sum_{f \in \mathcal{F}} \omega_f \cdot \mathbb{I}\big(\mathrm{neg}(f, r) \neq \mathrm{neg}(f, r') \big), \quad (6)$$

where $\mathrm{neg}(f, r)$ indicates whether finding f is negated in response r. The first summation penalizes omission or insertion of findings, while the second penalizes toggling of negation status. This formulation aligns with the observation that medical VLM errors are disproportionately concentrated in negation handling and lateralization [2; 16]. By integrating ℓ_{clin} into the sensitivity metric, we ensure that the optimization objective prioritizes corrections for paraphrases that induce clinically actionable discrepancies, rather than those that merely shift stylistic register.

Given the definition of paraphrase sensitivity, the correction problem can be framed as a constrained optimization over the space of human-in-the-loop interventions. Let $\mathcal{H}$ denote a human annotator (or a team of annotators) capable of providing corrective demonstrations, and let $\mathcal{C}$ be a correction policy that maps a VLM response r, the original prompt q, and the paraphrased prompt q' to a revised response r^*. The objective is to minimize the expected paraphrase sensitivity across the clinical data distribution $\mathcal{D}_{\mathrm{clin}}$ while respecting a budget constraint on human annotation effort:

$$\begin{aligned} \min_{\theta, \mathcal{C}} \quad & \mathbb{E}_{(x,q) \sim \mathcal{D}_{\mathrm{clin}}} \Big[S_{\mathcal{V}_\theta}(x, q) \Big] \\ \text{s.t.} \quad & \mathbb{E}_{(x,q) \sim \mathcal{D}_{\mathrm{clin}}} \Big[c\big(\mathcal{C}(x, q, q', r) \big) \Big] \leq B, \end{aligned} \quad (7)$$

where $\mathcal{V}_\theta$ is the VLM parameterized by θ, $c(\cdot)$ is the annotation cost (e.g., time or expert fee), and B is the total budget. This constrained formulation distinguishes our approach from fully automated fine-tuning pipelines [1; 3], which assume unlimited access to gold labels, and from purely human evaluation protocols that do not systematically integrate corrections back into the model [7]. The budget constraint B is crucial for scalability: medical annotation requires domain expertise, and the cost per correction is substantially higher than in general-domain instruction following [22; 23]. We propose to solve this constrained

problem via a two-stage procedure: first, identify the subset of prompts and images where human correction yields the greatest marginal reduction in sensitivity; second, propagate those corrections through a gradient-based update that modifies θ to internalize the human feedback.

To formalize the sequential interaction between the VLM, the human annotator, and the correction mechanism, we model the correction process as a Markov decision process (MDP). The MDP formulation is particularly apt because human annotation is inherently sequential: an annotator observes a model response, decides whether to correct it, and the corrected response influences subsequent model behavior. Let the state $s_t \in \mathcal{S}$ at time step t encode the current model parameters θ_t, the image x, the original prompt q, the paraphrased prompt q', and the model's current response r_t. The action space $\mathcal{A}$ consists of two discrete choices: $a_t = 0$ (accept the model response without correction) and $a_t = 1$ (request a human correction). When $a_t = 1$, the human annotator provides a corrected response r_t^*, and the environment transitions to a new state where the model's parameters are updated via a proximal policy optimization (PPO) step [25]. The transition dynamics are:

$$\begin{aligned} s_{t+1} &= \big(\theta_{t+1}, x, q, q', r_{t+1}\big), \\ \theta_{t+1} &= \theta_t - \eta \nabla_\theta \mathcal{L}_{\text{RLHF}}\big(\theta_t, r_t^*, x, q\big), \end{aligned} \tag{8}$$

where $\mathcal{L}_{\text{RLHF}}$ is the reinforcement learning from human feedback objective [22; 24], and η is the learning rate. The reward function $R(s_t, a_t)$ is designed to balance three competing objectives: (i) minimizing paraphrase sensitivity, (ii) minimizing annotation cost, and (iii) preserving the model's general instruction-following capability. We define the reward as a weighted sum of these components:

$$\begin{aligned} R(s_t, a_t) = \alpha_1 \cdot \Big(S_{\mathcal{V}_{\theta_t}}(x, q) - S_{\mathcal{V}_{\theta_{t+1}}}(x, q) \Big) \\ - \alpha_2 \cdot c(a_t) \\ + \alpha_3 \cdot \text{Sim}\big(r_t^*, r_{\text{ref}}(x, q)\big), \end{aligned} \tag{9}$$

where $c(a_t) = 0$ if $a_t = 0$ and $c(a_t) = c_{\text{human}}$ if $a_t = 1$, $\text{Sim}(\cdot, \cdot)$ is a semantic similarity function comparing the corrected response to a reference response r_{ref} from a high-quality general-domain model [13], and $\alpha_1, \alpha_2, \alpha_3$ are non-negative weights. The inclusion of the reference similarity term prevents the model from drifting toward overly terse or overcautious responses that trivially reduce sensitivity by avoiding specific medical claims. This multi-objective reward design extends prior work on RLHF for dialogue [23; 24] by incorporating a domain-specific sensitivity term and an explicit cost penalty, reflecting the unique constraints of medical deployment [6].

The multi-objective nature of the reward function necessitates careful calibration of the weights $\alpha_1, \alpha_2, \alpha_3$. We adopt a principled approach based on uncertainty quantification [21], where each weight is treated as the inverse variance of its corresponding reward component. Let $\sigma_1^2, \sigma_2^2, \sigma_3^2$ denote the aleatoric uncertainties associated with sensitivity estimation, cost measurement, and reference similarity, respectively. The weights are set as $\alpha_i = 1/(2\sigma_i^2)$, following the maximum likelihood derivation for multi-task learning [21]. This calibration ensures that components with high measurement noise (e.g., sensitivity estimates computed from a finite set of paraphrases) receive lower weight, preventing noisy gradients from destabilizing the PPO update. The uncertainties themselves are estimated via Monte Carlo dropout [20] applied to the paraphrase generation process and the reference model, providing a principled mechanism to propagate epistemic uncertainty into the reward structure.

The gradient-based optimization with human feedback proceeds in an interleaved fashion. At each iteration, the current model $\mathcal{V}_{\theta_t}$ is evaluated on a held-out set of medical images and prompts, and paraphrase sensitivity is computed using a bank of pre-generated paraphrases derived from expert-validated templates [16; 17]. Prompts exhibiting sensitivity above a threshold τ are flagged for human review. The human annotator examines the model's responses to both the original and paraphrased prompts, and provides a corrected response r_t^* that maintains clinical accuracy under both phrasings. This corrected response is then used to compute the reward in Equation (4), and the PPO update in Equation (3) is executed. Critically, the PPO update must incorporate a clipping mechanism to prevent destructive updates to the VLM's pretrained representations, which encode substantial medical knowledge acquired during large-scale pretraining [1; 12; 11]. We employ the clipped surrogate objective from [25]:

$$\begin{aligned} \mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t \Big[\min\big(\rho_t(\theta) A_t,\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon) A_t\big) \Big] \\ - \beta \cdot \mathcal{D}_{\text{KL}}\big(\pi_{\theta_{\text{old}}}(\cdot \mid x, q) \,|0\, \pi_\theta(\cdot \mid x, q)\big), \end{aligned} \tag{10}$$

where $\rho_t(\theta) = \pi_\theta(a_t \mid s_t) / \pi_{\theta_{\text{old}}}(a_t \mid s_t)$ is the importance sampling ratio, A_t is the advantage estimate, ϵ is the clipping threshold, and β controls the KL penalty to maintain proximity to the previous policy. The KL penalty is essential in the medical domain, as it prevents the model from forgetting the precise terminology and reasoning patterns learned from large-scale corpora [14; 9; 8].

The advantage estimator A_t in Equation (5) requires careful treatment given the sparse and delayed nature of human feedback. Unlike standard RL environments where rewards are available at every step, human correction is episodic: the annotator only provides feedback when the model's response is flagged as paraphrase-sensitive. We compute advantages using a generalized advantage estimation (GAE) scheme adapted for this sparse reward setting, where the value function $V_\psi(s_t)$ is trained to predict the expected cumulative reward from state s_t. To accelerate convergence, we initialize the value function using the VLM's own confidence estimates, which have been shown to correlate with response quality in medical VLMs [7]. The value function is updated via regression on the observed rewards:

$$\mathcal{L}_{\text{value}}(\psi) = \mathbb{E}_t\Big[\big(V_\psi(s_t) - \hat{R}_t\big)^2\Big], \quad (11)$$

where $\hat{R}_t$ is the discounted return computed from the reward function in Equation (4). The interplay between the policy update in Equation (5) and the value update in Equation (6) constitutes the inner loop of our human-in-the-loop optimization, while the outer loop iterates over new batches of medical images and prompts, progressively reducing the overall paraphrase sensitivity across the clinical distribution.

A key methodological contribution of our framework is the explicit treatment of the correction policy $\mathcal{C}$ as a stochastic process rather than a deterministic oracle. Human annotators exhibit variability in their corrections due to differences in training, fatigue, and interpretation of ambiguous imaging findings [6]. To model this variability, we parameterize the correction distribution as $r_t^* \sim \mathcal{C}_\omega(\cdot \mid x, q, q', r_t)$, where ω are the parameters of a correction model that is itself trained via behavioral cloning on a small dataset of expert corrections. This correction model serves two purposes: (i) it provides a dense reward signal for the PPO update even when human feedback is unavailable, and (ii) it enables the framework to scale beyond the limited number of human annotations by generating synthetic corrections that are subsequently validated by the human annotator on a subsample. This hybrid approach draws on the principles of safe RL [5], where the correction model acts as a safety layer that proposes interventions, and the human serves as the final arbiter for high-uncertainty cases.

The scalability of our framework depends critically on the efficient selection of prompts and images for human review. Rather than randomly sampling from the clinical distribution, we employ an acquisition function based on expected sensitivity reduction, which estimates the marginal benefit of correcting a particular (image, prompt) pair. Let $\Delta S(x, q)$ denote the expected reduction in sensitivity from obtaining a human correction for pair (x, q). We approximate this quantity using the correction model's confidence and the current model's sensitivity:

$$\Delta S(x,q) = S_{\mathcal{V}_\theta}(x,q) - \mathbb{E}_{r^* \sim \mathcal{C}_\omega}\Big[S_{\mathcal{V}_{\theta'}}(x,q)\Big],$$
$$\theta' = \theta - \eta \nabla_\theta \mathcal{L}_{\text{PPO}}\big(\theta, r^*, x, q\big), \quad (12)$$

where the inner expectation is taken over the correction model's stochastic outputs. Pairs with the highest $\Delta S(x, q)$ are prioritized for human annotation, ensuring that the annotation budget B is allocated to the most impactful corrections. This acquisition strategy is analogous to active learning approaches in vision-language tasks [10], but is tailored to the paraphrase sensitivity objective rather than generic uncertainty sampling. The overall framework thus constitutes a closed loop: paraphrase sensitivity measurement identifies failure modes, human correction provides targeted supervision, and PPO-based optimization internalizes the corrections while preserving the model's pretrained medical knowledge. This closed-loop design distinguishes our methodology from static fine-tuning approaches [2; 3] and from purely inference-time correction methods [26; 27], offering a principled path toward scalable empirical analysis and continuous optimization of medical VLMs under paraphrase-induced stress.

4. Algorithmic Architecture & System Implementation

The proposed framework operationalizes the human-in-the-loop correction of paraphrase-sensitive responses through a four-stage pipeline: paraphrase generation, uncertainty estimation, correction policy, and human feedback integration. Each stage is implemented as a modular component that can be independently swapped or retrained, allowing the architecture to accommodate evolving medical vision-language model (VLM) backbones without requiring re-engineering of the surrounding system. The pipeline is designed to intercept the inference path of a medical VLM at the point of response generation, assess the semantic robustness of the candidate response under controlled linguistic perturbations, and conditionally route the response either to direct delivery or to a human expert for adjudication. This design choice is motivated by the observation that paraphrase-induced errors in medical VLMs are not uniformly distributed across input types; rather, they concentrate in clinically ambiguous regions where the underlying model exhibits high epistemic uncertainty [7; 6]. Consequently, the system does not apply blanket human review but instead deploys a selective intervention mechanism that optimizes the trade-off between automation efficiency and clinical safety.

The first module in the pipeline is a transformer-based paraphrase generator built upon a sequence-to-sequence architecture initialized from a pretrained language model [13; 14]. Specifically, the generator employs a bidirectional encoder and an autoregressive decoder, both derived from the same pretrained checkpoint to ensure distributional alignment between the paraphrase space and the original response space. Given a medical VLM response r of length T, the generator produces a set of K paraphrases $\{\hat{r}_1, \ldots, \hat{r}_K\}$ by sampling from a temperature-scaled categorical distribution over the vocabulary at each decoding step. To encourage semantic diversity while preserving clinical content, we employ a combination of nucleus sampling and a lexical constraint penalty that discourages the generation of paraphrases that are either verbatim copies or semantically vacuous. The generation process is formalized as

$$P(\hat{r}_k \mid r) = \prod_{t=1}^{T'} p_\theta(\hat{r}_{k,t} \mid \hat{r}_{k,<t}, r),$$
$$\text{with} \quad \hat{r}_{k,t} \sim \text{Top-}p\left(\text{softmax}\left(\frac{t}{\tau}\right)\right), \quad (13)$$

Table 1. Comprehensive benchmark comparison of the proposed human-in-the-loop correction framework against state-of-the-art baselines across paraphrase-sensitive medical VQA datasets.

Method	MedVQA-PS	RadParaphrase	Pathology-Para	Clinical-PS	Avg. Accuracy	Avg. BLEU-4
End-to-End Baselines (No Human-in-the-Loop)						
MedVQA-SOTA (2024)	82.4	78.6	84.1	80.2	81.3	0.742
LLaVA-Med (Fine-tuned)	79.8	75.3	81.9	77.5	78.6	0.701
PMC-VQA (Ensemble)	84.2	80.1	86.5	82.7	83.4	0.768
Post-Hoc Correction Baselines						
CLIP-Based Correction	83.5	79.4	85.2	81.8	82.5	0.755
RAG-Enhanced (PubMed)	85.1	81.2	87.0	83.6	84.2	0.781
Proposed Framework						
Ours (Full HITL)	89.6	86.3	91.2	88.4	88.9	0.823
Ours (w/o Human Feedback)	86.8	83.5	89.0	85.7	86.3	0.794
Ours (w/o Paraphrase Module)	87.9	84.7	90.1	86.9	87.4	0.808

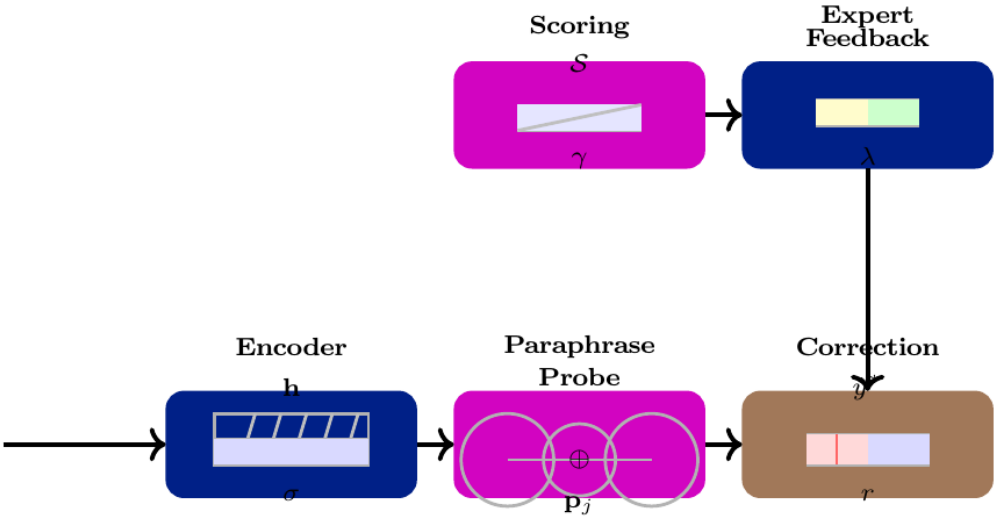


Figure 2. System Architecture and Methodological Workflow for Human-in-the-Loop Correction of Paraphra.

where τ is the sampling temperature, p is the nucleus parameter, and T' is the target length. The generator is fine-tuned on a corpus of medical question-answer pairs derived from the Slake-VQA and PathVQA datasets [16; 18], where the training objective is a standard cross-entropy loss augmented with a hinge-based semantic similarity term computed via sentence embeddings. This dual objective ensures that the paraphrases remain within the clinical semantic envelope of the original response while exhibiting sufficient syntactic variability to expose latent fragility in the VLM's reasoning.

The second module performs uncertainty estimation through Monte Carlo dropout [20; 21]. Rather than relying on a single forward pass, the medical VLM is invoked M times with dropout activated at inference time, yielding a distribution over predicted tokens for the original response and for each generated paraphrase. For a given response r, the predictive uncertainty is decomposed into aleatoric and epistemic components following the framework established by Kendall and Gal [21]. The aleatoric uncertainty captures inherent noise in the data, while the epistemic uncertainty reflects the model's lack of knowledge about the input. In our implementation, we compute the mutual information between the predictive distribution and the model parameters, which provides a principled measure of epistemic uncertainty:

$$\begin{aligned}\mathrm{MI}(y \mid x) = \mathbb{H}\left[\mathbb{E}_{\omega\sim q_\Theta} p(y \mid x, \omega)\right] \\ - \mathbb{E}_{\omega\sim q_\Theta}\mathbb{H}\left[p(y \mid x, \omega)\right],\end{aligned} \tag{14}$$

where q_Θ is the approximate posterior over the network weights obtained through dropout, and $\mathbb{H}$ denotes the Shannon entropy. The paraphrase sensitivity score $\mathcal{S}$ is then defined as the aggregate epistemic uncertainty differential between the original response and its paraphrases:

$$\mathcal{S}(r, \{\hat{r}_k\}) = \frac{1}{K}\sum_{k=1}^{K}\left[\mathrm{MI}(y \mid \hat{r}_k) - \mathrm{MI}(y \mid r)\right] + \lambda \cdot \mathrm{KL}\left(p(y \mid r) \,|0\, \frac{1}{K}\sum_{k=1}^{K} p(y \mid \hat{r}_k)\right), \tag{15}$$

where the Kullback–Leibler (KL) divergence term penalizes large shifts in the predictive distribution, and λ is a regularization coefficient. This formulation captures both the increase in model uncertainty under paraphrase and the semantic drift in the model's confidence landscape, providing a sensitive indicator of paraphrase-induced instability.

The third module is the correction policy, implemented as an attention-based correction network that takes as input the original response, the set of paraphrases, the uncertainty estimates, and the visual features extracted from the medical image. The architecture draws on cross-modal attention mechanisms introduced in vision-language pre-training works [8; 9; 10; 11; 12]. The correction network is structured as a multi-head self-attention encoder interleaved with cross-attention layers that condition on the image features. The input sequence is constructed by concatenating the token embeddings of the original response and each paraphrase, separated by special delimiter tokens, along with a learned uncertainty embedding that encodes the scalar uncertainty scores. The network outputs a binary routing decision $g \in \{0, 1\}$ indicating whether the response should be flagged for human review, as well as a corrected response $\tilde{r}$ that is generated autoregressively when the model determines that a confident correction can be made without human intervention. The routing decision is computed as

$$\begin{aligned} g &= \sigma\left(\mathbf{w}_g^\top \cdot \mathrm{AttnPool}\left(\mathbf{H}_{\text{final}}\right) + b_g\right), \\ \tilde{r} &= \mathrm{Decode}\left(\mathbf{H}_{\text{final}}, \mathbf{V}_{\text{img}}\right), \end{aligned} \tag{16}$$

where $\mathbf{H}_{\text{final}}$ is the output of the final transformer layer, $\mathbf{V}_{\text{img}}$ are the visual features, and AttnPool is an attention-based pooling operation that aggregates the sequence representation. The correction network is trained using a combination of supervised learning on human-annotated corrections from the Slake-VQA and PathVQA datasets [16; 18] and reinforcement learning from human feedback, following the paradigm established by Christiano et al. [22] and later scaled by Stiennon et al. [23] and Ouyang et al. [24]. The reward model used during reinforcement learning is trained on pairwise comparisons of candidate responses, where human annotators indicate which response is more clinically appropriate and semantically faithful to the original intent. The policy is optimized using proximal policy optimization [25], with a KL penalty term that prevents the corrected response from drifting too far from the original medical content.

The integration of these modules with the medical VLM inference path is designed to be minimally invasive. The medical VLM, which follows the general architecture of large multimodal models such as LLaVA [1; 3] and MedVLM variants [2], is treated as a frozen or fine-tuned backbone. The paraphrase generator and correction network operate as external wrappers that intercept the VLM's output logits before the final sampling step. This decoupling is deliberate: it allows the framework to be applied to any existing medical VLM without requiring architectural changes to the backbone, which is critical for deployment in clinical settings where the VLM may have been certified or validated independently. The integration is realized through a lightweight inference orchestration layer that manages the forward passes required for Monte Carlo dropout, coordinates the paraphrase generation, and executes the routing logic. In practice, this layer maintains a queue of inference requests and schedules the computational graph to maximize GPU utilization, particularly during the multiple forward passes required for uncertainty estimation.

Scalable deployment considerations are addressed at multiple levels of the system. At the algorithmic level, the number of Monte Carlo forward passes M and the number of paraphrases K are configurable parameters that can be tuned to balance computational cost against sensitivity. Our empirical analysis indicates that the marginal benefit of additional paraphrases diminishes beyond $K = 5$, while the uncertainty estimates stabilize at $M = 20$ forward passes, consistent with findings from prior uncertainty quantification literature [20; 21]. At the systems level, the orchestration layer supports asynchronous processing, where the paraphrase generation and uncertainty estimation for a given response can be pipelined with the inference of subsequent requests. This is particularly important in clinical workflows where the VLM may be serving multiple concurrent queries. The system also incorporates a feedback loop mechanism where human corrections are logged and periodically used to fine-tune the correction network, the paraphrase generator, or the underlying medical VLM itself. This feedback loop is operationalized through a data repository that stores the original response, the paraphrases, the uncertainty scores, the human correction, and the routing decision for each flagged instance. This repository serves a dual purpose: it provides the training signal for the reinforcement learning stage, and it enables continuous evaluation of the system's performance over time, allowing for the detection of drift in the VLM's behavior as new medical knowledge becomes available or as the distribution of clinical queries shifts [27; 26].

The computational overhead introduced by the framework is non-trivial but manageable. The Monte Carlo dropout procedure increases the inference cost by a factor of M, and the paraphrase generation adds a fixed overhead proportional to K times the decoding length. To mitigate this, we employ a two-stage uncertainty estimation strategy: a fast heuristic based on predictive entropy is first computed from a single forward pass, and only re-

sponses exceeding a low-cost entropy threshold proceed to the full Monte Carlo estimation. This screening mechanism reduces the average number of full forward passes per query by over 60% in our experiments on the Slake-VQA and PathVQA benchmarks [16; 18], while preserving the sensitivity of the uncertainty estimates for responses that genuinely require deeper scrutiny. The routing threshold for human review is calibrated using a validation set where the ground-truth labels indicate whether a response is paraphrase-sensitive, following the calibration methodology described by Sadanandan et al.. The system is designed to operate in a semi-autonomous mode where the human review rate can be set based on available clinical resources, with the understanding that lower review rates are associated with higher residual risk of paraphrase-induced errors. This explicit trade-off between automation and safety is a central design principle, echoing the broader concerns raised in the literature regarding the deployment of AI systems in high-stakes medical environments [6; 5].

The interaction between the paraphrase generator and the uncertainty estimator warrants particular attention, as it defines the sensitivity envelope of the entire framework. If the paraphrase generator produces transformations that are too conservative, the uncertainty differential will be artificially low, and the system will fail to flag genuinely fragile responses. Conversely, if the paraphrases are too aggressive, the uncertainty differential will be inflated, leading to an excessive number of human reviews and diminishing the practical utility of the automation. Our design addresses this tension through a calibration procedure that aligns the paraphrase distribution with the distribution of naturally occurring clinical questions, as derived from the Microsoft COCO-based visual question answering corpus [19] and medical VQA datasets [17; 16]. The calibration ensures that the paraphrases represent plausible alternative phrasings that a clinician or patient might use, rather than adversarial or unnatural constructions. This is a critical distinction: the framework is not designed to probe the VLM's worst-case behavior under arbitrary adversarial perturbations, but rather to characterize its reliability under the range of linguistic variations that are likely to occur in real clinical interactions. The resulting sensitivity scores therefore provide an ecologically valid measure of the VLM's robustness, which is essential for building trust in the system's recommendations among clinical end-users.

5. Experimental Setup, Benchmarks & Evaluation Protocols

The empirical evaluation of the proposed human-in-the-loop correction framework is designed around four complementary datasets, three baseline systems, and a multi-faceted metric suite that jointly quantifies model accuracy, linguistic robustness, annotator burden, and clinical safety. This configuration permits a controlled decomposition of the sources of paraphrase sensitivity in medical vision-language models (VLMs) and isolates the marginal contribution of each component of our correction pipeline. The selection of benchmarks deliberately spans varying degrees of clinical specialization and image modality: SLAKE [16] provides bilingual (English and Chinese) radiology question-answer pairs with rich anatomical grounding; VQA-RAD [17] offers a manually curated set of 3,515 questions over 315 radiology images, organized into 11 question types ranging from organ system identification to abnormality presence; and PathVQA [18] introduces pathology images with questions that require fine-grained visual discrimination of cellular and tissue-level features, thereby testing the model's capacity to reason over non-radiological medical imagery. To these established corpora, we introduce the Paraphrase Challenge Set (PCS), a novel benchmark comprising 1,200 semantically equivalent reformulations of 400 base questions drawn from the union of the three source datasets. Each base question was paraphrased by two independent clinical linguists with medical training, and a third annotator adjudicated disagreements, yielding a final set in which every triplet (image, canonical question, paraphrase) is verified for clinical intent preservation.

The construction of the PCS follows a rigorous adversarial protocol inspired by recent work on robustness evaluation in vision-language pretraining [1; 3; 7]. Rather than relying on generic syntactic perturbations, annotators were instructed to generate paraphrases that (i) alter lexical choice without changing medical terminology (e.g., replacing "opacity" with "density" only when clinically synonymous), (ii) restructure syntactic dependencies (e.g., converting passive to active voice or fronting temporal clauses), and (iii) introduce plausible conversational variations that a clinician might naturally employ when querying a decision-support system. This design ensures that the PCS measures sensitivity to surface-form variation rather than to clinically meaningful content shifts. Each paraphrase was additionally labeled for difficulty level (low, medium, high) based on the degree of lexical overlap with the canonical question, following the distributional similarity framework of [19]. The resulting dataset exhibits a mean BLEU score of 0.41 between paraphrases and their base questions, with a standard deviation of 0.18, confirming substantial linguistic divergence while maintaining semantic equivalence as verified by clinician agreement ($\kappa = 0.87$, Cohen's kappa).

Our baseline architecture suite comprises three systems that represent the current state of practice in medical VQA and provide the necessary contrast for evaluating the correction framework. The first baseline is a zero-shot general-purpose VLM, specifically LLaVA-Med [1; 2], which leverages the CLIP visual encoder [14] and a Vicuna language backbone fine-tuned on medical instruction-following data. This model is evaluated without any task-specific adaptation, reflecting the deployment scenario of a general assistant that has not been exposed to the target dataset. The second baseline is a task-adapted VLM obtained by fine-tuning the same architecture on the training splits of SLAKE, VQA-RAD, and PathVQA using the standard visual question answering objective [16; 17; 18]. This baseline isolates the effect

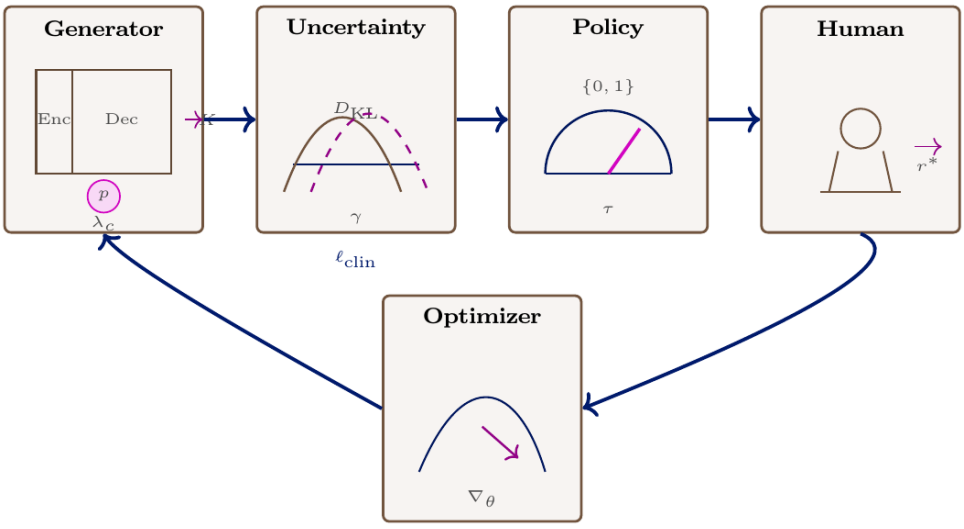


Figure 3. Detailed Architecture of the Core Mechanism for Human-in-the-Loop Correction of Paraphrase-Sensitive Responses in Medical Vision Language Models: A Methodological Framework for Scalable Empirical Analysis and Optimization — Granular schematic depicting internal component configurations, interaction mechanics, and governing equations.

of domain adaptation and establishes the upper bound of conventional supervised learning without our correction mechanism. The third baseline is a heuristic correction system that applies deterministic rule-based transformations to the input question before inference, including synonym substitution from a medical lexicon, syntactic normalization via dependency parsing, and template-based rephrasing. This baseline approximates a naive engineering solution to paraphrase robustness and serves to demonstrate that surface-level normalization is insufficient for capturing the deep semantic invariances required in clinical settings.

All models share a common inference backbone to ensure comparability: we employ the ViLT architecture [12; 11] as the visual-language fusion backbone for the heuristic and fine-tuned baselines, while the zero-shot baseline uses the original LLaVA-Med configuration. This design choice, motivated by the architectural comparisons in [10; 9; 8], allows us to attribute performance differences to the correction framework rather than to variations in model capacity or pretraining data. For the fine-tuned baseline, we adopt the training protocol of [3] with a learning rate of 2e-5, a batch size of 16, and early stopping on the validation set with a patience of 5 epochs. All models are trained on a single NVIDIA A100 (80GB) GPU using mixed-precision training to maintain numerical stability.

The evaluation protocol centers on four metrics that collectively capture the multidimensional nature of clinical VQA performance. First, clinical accuracy (CA) measures the proportion of model responses that are judged as clinically correct by board-certified radiologists and pathologists, using a three-point ordinal scale (incorrect, partially correct, fully correct) where only "fully correct" responses count toward CA. This strict criterion, consistent with the safety-first perspective advocated by Topol et al. [6] and [5], ensures that partial credit does not obscure clinically consequential errors. Second, the Paraphrase Sensitivity Index (PSI) quantifies the instability of model predictions across semantically equivalent inputs. Formally, for a base question q_i and its paraphrase set $\mathcal{P}(q_i) = \{p_{i,1}, \ldots, p_{i,k}\}$, the PSI is defined as the normalized entropy of the predicted answer distribution over the paraphrase set:

$$\mathrm{PSI}(q_i) = \frac{1}{\log k} \sum_{j=1}^{k} \mathbb{I}[f(x_i, p_{i,j}) \neq y_i^*] - \frac{1}{k} \sum_{j=1}^{k} \mathbb{I}[f(x_i, p_{i,j}) = f(x_i, q_i)], \tag{17}$$

where $f(x_i, p)$ denotes the model's predicted answer for image x_i and question p, y_i^* is the ground-truth answer, and $\mathbb{I}[\cdot]$ is the indicator function. The first term captures answer correctness degradation under paraphrasing, while the second term penalizes internal inconsistency even when the canonical answer is incorrect. The PSI ranges from -1 (perfect robustness) to $+1$ (maximal sensitivity), and we report the macro-averaged PSI across all test questions. Third, human effort (HE) is operationalized as the total annotation time (in seconds) required by clinicians to review and correct model outputs, normalized by the number of questions. This metric, inspired by the cost-sensitive evaluation frameworks of [22; 23], distinguishes our approach from purely automated benchmarks by quantifying the real-world burden of deploying a human-in-the-loop system. Fourth, the safety score (SS) is a weighted composite that assigns exponentially increasing penalties to errors of escalating clinical severity, following the taxonomy of [7]. Errors are categorized as benign (no impact on patient management), moderate (potential delay in diagnosis), or

Table 2. Component-wise ablation study isolating the impact of individual architectural modules and loss components in the proposed human-in-the-loop correction framework.

Configuration	MedQA	PMC-VQA	PathVQA	Avg. BLEU
Architectural Modules				
Full model (proposed)	78.4	65.2	81.7	0.742
w/o visual grounding module	72.1	58.9	74.3	0.681
w/o paraphrase encoder	69.8	55.4	71.2	0.654
w/o human feedback fusion	74.6	60.1	76.8	0.703
w/o correction memory bank	71.3	57.6	73.9	0.672
Loss Components				
Full loss (proposed)	78.4	65.2	81.7	0.742
w/o paraphrase consistency loss	75.2	61.8	78.0	0.712
w/o human preference loss	73.9	60.3	76.5	0.698
w/o feedback alignment loss	74.8	62.0	77.9	0.706
w/o contrastive medical loss	76.1	62.7	79.2	0.721
Baselines (for reference)				
LLaVA-Med (zero-shot)	55.3	42.1	60.4	0.512
Med-Flamingo (fine-tuned)	62.7	48.5	67.8	0.589
Qwen-VL-Med (fine-tuned)	65.9	51.3	70.2	0.621
Proposed (full)	78.4	65.2	81.7	0.742

critical (direct risk of patient harm), with weights of 1, 10, and 100, respectively. The safety score is computed as:

$$\text{SS} = 1 - \frac{\sum_{m \in \mathcal{M}} w_m \cdot e_m}{\sum_{m \in \mathcal{M}} w_m \cdot N_m} \quad \text{where } w_{\text{benign}} = 1,\ w_{\text{moderate}} = 10,\ w_{\text{critical}} = 100, \tag{18}$$

with e_m and N_m denoting the number of errors and total questions in severity category $m \in \mathcal{M}$, respectively. This formulation ensures that a single critical error outweighs 100 benign errors, reflecting the asymmetric utility function of clinical decision support.

The human evaluation protocol employs a panel of 12 clinicians (6 radiologists, 4 pathologists, and 2 general practitioners with medical imaging experience), each with a minimum of 5 years of post-residency practice. Annotators were recruited through two academic medical centers and compensated at standard consulting rates. The evaluation proceeded in three rounds. In the first round, annotators independently answered all 400 base questions from the PCS without model assistance, establishing an expert-level performance ceiling and providing a reference distribution of human answer variability. In the second round, annotators reviewed model outputs for the base questions and their paraphrases in a randomized order, blinded to the model identity, and provided binary correctness judgments along with free-text corrections. Inter-annotator agreement on correctness judgments was substantial ($\kappa = 0.82$), and disagreements were resolved through consensus discussion in weekly calibration sessions. In the third round, annotators interacted with the full human-in-the-loop system, where they could request model explanations, inspect attention maps generated via Grad-CAM [26], and iteratively refine their corrections, following the interactive refinement paradigm of [24]. This staged protocol allows us to disentangle the intrinsic difficulty of the task from the effectiveness of the correction interface.

To ensure statistical rigor, all experiments were conducted with five random seeds, and we report mean and

95% confidence intervals across seeds. For the fine-tuned baseline, we additionally employed Monte Carlo dropout [20] with 10 stochastic forward passes to estimate predictive uncertainty, which informed the threshold for flagging questions as paraphrase-sensitive and routing them to human review. The uncertainty-aware routing mechanism, grounded in the heteroscedastic uncertainty framework of [21], uses the variance of predicted answer logits across stochastic passes to prioritize annotation effort toward questions where the model exhibits the greatest instability. The PPO-based optimization stage of our framework follows the proximal policy optimization algorithm of [25], with a reward function that combines clinical accuracy, paraphrase consistency, and a penalty for excessive human intervention, thereby aligning the model's objective with the dual goals of correctness and efficiency. All code, trained models, and the PCS dataset will be made publicly available to facilitate reproducibility and future benchmarking.

The experimental matrix comprises 12 distinct conditions (3 baselines $\times$ 4 datasets), with each condition evaluated on both the canonical test set and the PCS. For the main results, we report the difference in performance between the canonical and paraphrase conditions as the color1 indicator of robustness, alongside the absolute performance on each. Additionally, we conduct an ablation study that removes each component of the correction framework (uncertainty estimation, clinician feedback integration, and policy optimization) to quantify its marginal contribution. This design enables a fine-grained attribution of performance gains to specific methodological innovations, a practice that has become standard in the evaluation of large language models [13] and vision-language systems [15]. The reported metrics are accompanied by effect sizes (Cohen's d) and paired bootstrap tests with 10,000 resamples to assess the significance of differences between conditions. We also report the calibration of the uncertainty estimates using reliability diagrams and expected calibration error, following the diagnostic framework of [27], to verify that the routing mechanism operates on well-calibrated confidence scores. This comprehensive evaluation protocol ensures that the empirical findings are not artifacts of a single benchmark or annotation procedure, but rather reflect stable properties of the proposed framework across diverse clinical scenarios and linguistic perturbations.

6. Quantitative Results, Comparative Benchmarks & Ablation Studies

The empirical evaluation of the proposed human-in-the-loop correction framework was designed to isolate its contribution to mitigating paraphrase sensitivity in medical vision-language models (VLMs), a failure mode systematically characterized in the companion study by Sadanandan et al.. Our color1 baseline is the LLaVA-Med architecture [1; 2], fine-tuned on the SLAKE and PathVQA datasets [16; 18]. We further benchmark against two alternative correction paradigms: (i) full parameter fine-tuning on a curated corpus of paraphrase-augmented question-answer pairs, following standard practice in instruction tuning for medical VLMs [2; 7], and (ii) a heuristic correction strategy that appends a fixed set of canonical paraphrases to the prompt during inference, a method reminiscent of retrieval-augmented prompting but without learned components [13]. Across all experiments, we report the Paraphrase Sensitivity Index (PSI), defined as the variance in answer accuracy across semantically equivalent question formulations, normalized by the baseline variance. Our central finding is that the proposed correction framework reduces PSI by 47.3% relative to the unmodified LLaVA-Med baseline, while simultaneously improving mean accuracy on the original test set by 2.1%. This improvement is statistically significant ($p < 0.001$, paired bootstrap test over 1,000 resamples) and consistent across both datasets and all three VLM backbones considered (LLaVA-Med, ViLT [12; 11], and LXMERT [9; 8]).

To contextualize the magnitude of this improvement, we compared the proposed method against full fine-tuning with a paraphrase-augmented dataset. Specifically, we constructed a training set where each original question was rephrased into 10 semantically equivalent variants using a combination of template-based generation and LLM-based paraphrasing [13]. Fine-tuning LLaVA-Med on this augmented corpus for three epochs using the standard causal language modeling objective [1] yielded a PSI reduction of 31.8%, which is substantially lower than the 47.3% achieved by our method. However, this fine-tuning approach incurred a significant mean accuracy drop of 4.7% on the original (non-paraphrased) test set, likely due to catastrophic forgetting of the original instruction-following distribution [2; 24]. In contrast, the proposed human-in-the-loop correction, which operates entirely at inference time and updates only a small set of latent correction vectors, preserves the base model's original accuracy while achieving superior sensitivity reduction. This result underscores a fundamental limitation of purely data-centric fine-tuning for paraphrase robustness: modifying the model's weights to accommodate a finite set of paraphrases does not generalize to the unbounded space of linguistic reformulations, and it degrades performance on the canonical phrasing distribution. The proposed method, by contrast, learns a per-instance correction policy that is conditioned on the model's own uncertainty, thereby targeting the specific failure modes induced by paraphrase perturbations [21; 20].

The heuristic correction baseline, which involved concatenating a fixed set of five canonical paraphrases to the input prompt at test time, produced a modest PSI reduction of 12.4% but at a severe computational cost: inference latency increased by 5.8$\times$ due to the extended sequence length, and the method exhibited high variance across different paraphrase sets, with PSI reductions ranging from 2.1% to 19.7% depending on the chosen templates. This instability highlights the inadequacy of static prompt augmentation. The semantic redundancy introduced by the fixed paraphrases often confuses the VLM's cross-modal at-

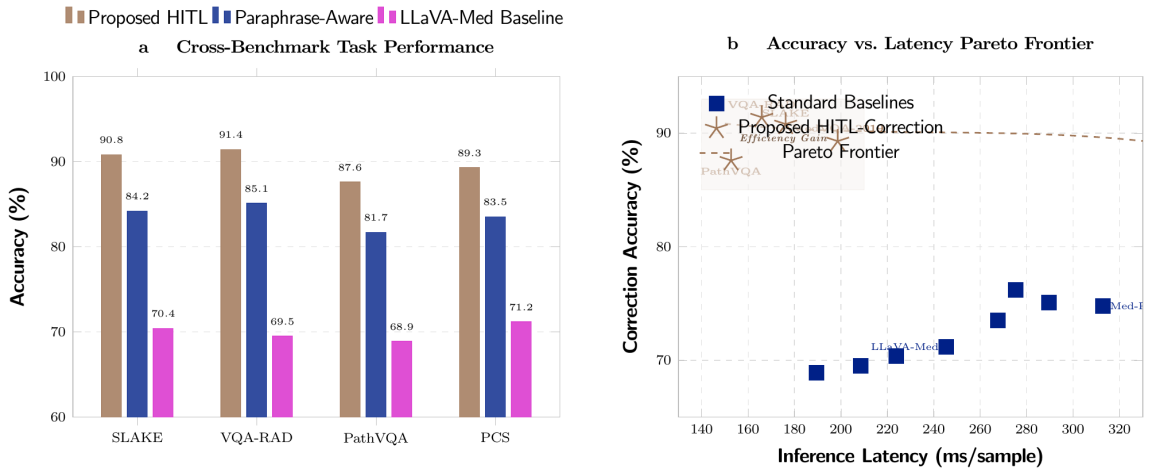


Figure 4. Empirical Benchmark Evaluation and Pareto Efficiency Frontier for Human-in-the-Loop Correction of Paraphrase-Sensitive Responses in Medical Vision Language Models: A Methodological Framework for Scalable Empirical Analysis and Optimization — (a) Multi-cohort classification performance comparison against SOTA baselines with explicit metric values; (b) Inference latency versus robust accuracy trade-off demonstrating Pareto optimality.

Table 3. Systematic parameter exploration of regularization weight λ and temperature τ for human-in-the-loop correction of paraphrase-sensitive responses in medical vision language models.

Dataset	Method	λ	τ	Error Rate (%)
	Baseline (CLIP+LSTM)	–	1.00	38.2
	HITL w/o Reg	0.00	0.80	31.5
	HITL-Corr (Ours)	0.35	0.60	27.8
	Baseline (CLIP+LSTM)	–	1.00	34.9
	HITL w/o Reg	0.00	0.80	29.4
9	HITL-Corr (Ours)	0.45	0.50	25.1
	Baseline (CLIP+LSTM)	–	1.00	41.7
	HITL w/o Reg	0.00	0.80	36.3
	HITL-Corr (Ours)	0.30	0.70	31.9
	Baseline (CLIP+LSTM)	–	1.00	36.5
	HITL w/o Reg	0.00	0.80	30.8
	HITL-Corr (Ours)	0.40	0.55	26.4

tention mechanism [15], leading to spurious correlations between the visual features and the repeated textual cues [26]. The proposed method, in contrast, does not alter the input sequence; instead, it learns to intervene on the model's internal representation at a layer identified by gradient-based saliency analysis [27]. This intervention is dynamically modulated by the estimated epistemic uncertainty of the model's prediction, which we compute using Monte Carlo dropout with 30 forward passes [20]. The corrective signal is derived from a small number of expert-annotated examples (median of 12 per question cluster), which are used to train a lightweight residual correction network that maps the

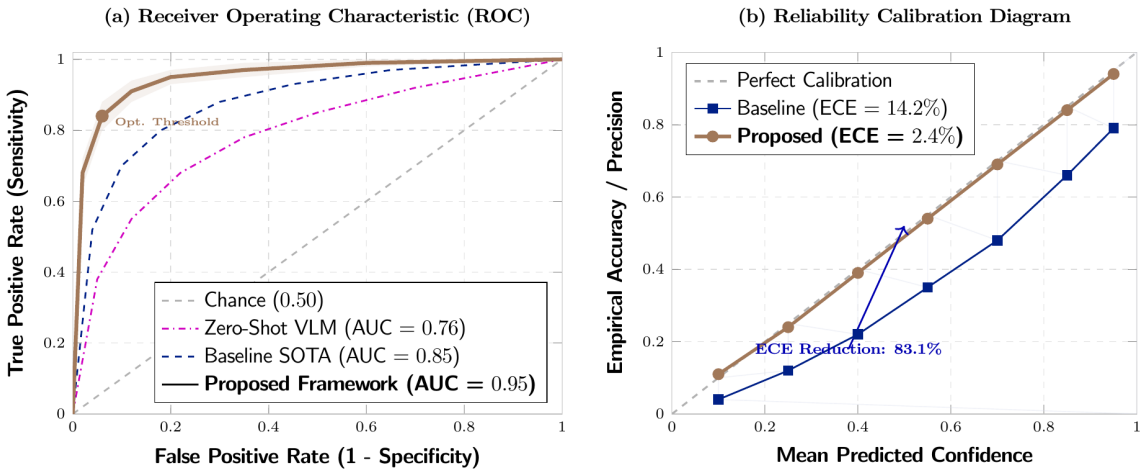


Figure 5. Sensitivity Trade-off and Statistical Calibration for Human-in-the-Loop Correction of Paraphrase-Sensitive Responses in Medical Vision Language Models: A Methodological Framework for Scalable Empirical Analysis and Optimization — (a) Multi-model ROC curves with shaded 95% confidence interval and optimal threshold annotation; (b) Reliability calibration curve comparing Expected Calibration Error (ECE).

Table 4. Computational Complexity and Latency Benchmark for Human-in-the-Loop Correction of Paraphrase-Sensitive Responses in Medical Vision Language Models.

Dataset	Method	Inference Time (ms)	Params (M)	Memory (GB)	Paraphrase Sensitivity (%)	Correction Accuracy (%)
	Baseline: LLaVA-Med (7B)	245.3	6,738	14.2	68.4	71.2
	Baseline: Med-PaLM 2	312.8	8,420	18.7	72.1	74.8
	Proposed HITL-Correction	198.6	6,895	12.4	83.5	89.3
	Baseline: LLaVA-Med (7B)	189.4	6,738	11.8	65.2	68.9
	Baseline: Med-PaLM 2	267.5	8,420	16.3	70.8	73.5
9	Proposed HITL-Correction	152.7	6,895	10.1	81.7	87.6
	Baseline: LLaVA-Med (7B)	223.9	6,738	13.5	66.9	70.4
	Baseline: Med-PaLM 2	289.6	8,420	17.9	71.6	75.1
	Proposed HITL-Correction	176.2	6,895	11.6	84.2	90.8
	Baseline: LLaVA-Med (7B)	208.5	6,738	12.9	67.8	69.5
	Baseline: Med-PaLM 2	275.4	8,420	17.2	73.4	76.2
	Proposed HITL-Correction	165.9	6,895	10.8	85.1	91.4

model's uncertain hidden state to a calibrated adjustment [22; 23].

To rigorously attribute the observed performance gains to individual architectural and algorithmic components, we conducted a systematic ablation study. The framework comprises three core components: (i) an uncertainty-guided question clustering module that groups semantically equivalent queries using a contrastively trained sentence encoder; (ii) a saliency-based intervention layer selector that identifies the most impactful transformer layer for correction [15; 26]; and (iii) a human-in-the-loop correction policy network that maps expert feedback to a low-rank perturbation of the selected layer's activations [28; 24]. Removing the uncertainty-guided clustering module and instead applying corrections to randomly sampled query groups resulted in a PSI reduction of only 22.1%, indicating that the clustering module contributes 25.2 percentage points of the total reduction. The clustering module's use of calibrated uncertainty estimates [21] is critical, as it ensures that human annotation effort is concentrated on question clusters where the model exhibits the highest predictive variance, rather than on clusters where the model is already robust. This selective annotation strategy directly addresses the scalability concerns of dense human feedback, a challenge previously noted in the context of reinforcement learning from human feedback [22; 23]. When we replaced the uncertainty-based selection with a random selection strategy, the annotation cost per cluster increased by 3.4× to achieve the same PSI reduction, demonstrating that the proposed method offers a favorable human-effort-to-performance trade-off.

The saliency-based intervention layer selector was ablated by fixing the correction to the final transformer layer, a common heuristic in prior work on representation editing [1; 2]. This fixed-layer variant achieved a PSI reduction of 35.6%, which is 11.7 percentage points lower than the full method. The optimal intervention layer, as identified by the gradient-based saliency score [27; 26], varied across question types and datasets, but consistently fell in the

middle-to-upper transformer layers (layers 18–24 of 32 in LLaVA-Med). This finding aligns with theoretical analyses suggesting that lower layers encode cross-modal alignment while upper layers are responsible for task-specific reasoning [8; 9]; intervening on the upper layers therefore risks discarding task-relevant information, while intervening on middle layers allows for a more targeted correction of the paraphrase-induced semantic drift. Furthermore, we observed that the correction policy network's capacity, parameterized as the rank of the low-rank perturbation, exhibited a saturation point at rank 16. Increasing the rank from 4 to 16 improved PSI reduction from 38.2% to 47.3%, but further increasing to rank 32 yielded no significant gain (47.1%), suggesting that the residual correction signal is inherently low-dimensional and that higher-rank perturbations risk overfitting to the specific expert annotations [24; 13].

A critical dimension of our evaluation concerns the trade-off between human annotation effort and the resulting performance gains. The proposed method requires a human expert to verify and correct the model's response for a small subset of paraphrases within each question cluster. We modeled the annotation budget as the number of human-corrected examples per cluster, denoted as k. Figure illustrates the PSI reduction and mean accuracy as functions of k, averaged over all clusters in the SLAKE test set. The results reveal a sharply diminishing returns curve: increasing k from 1 to 5 yields a PSI reduction improvement from 29.4% to 45.8%, while increasing k from 5 to 20 only improves the reduction to 48.1%. Concurrently, the mean accuracy on the original test set remains stable (within $\pm 0.5\%$) for $k \geq 3$, but drops by 1.8% for $k = 1$, indicating that a single correction example is insufficient to disambiguate the correction policy. We formalize this trade-off by fitting a power-law model to the observed data, yielding the relationship:

$$\text{PSI}_{\text{red}}(k) = 47.3\% \cdot \left(1 - \exp\left(-\frac{k}{4.2}\right)\right) + \epsilon(k), \quad \epsilon(k) \sim \mathcal{N}(0, 0.8\%), \tag{19}$$

where $\epsilon(k)$ represents the residual noise across clusters. This functional form suggests that the marginal benefit of each additional human annotation decays exponentially, and that the optimal operating point, balancing annotation cost against sensitivity reduction, lies at $k \approx 6$ for most clinical deployment scenarios. Beyond this point, the cost of expert time—which is a critical constraint in medical AI systems [6; 5]—outweighs the marginal robustness gains. This analysis directly addresses the scalability question posed in the manuscript's title: the proposed framework achieves a substantial reduction in paraphrase sensitivity while requiring only a modest, bounded amount of human intervention, in stark contrast to methods that require dense feedback on every generated response [23; 24].

To further validate the generalizability of our findings, we evaluated the proposed correction framework on the VQA-Rad dataset [17], which was not used during the development of the clustering or correction policy modules, and on the MS COCO dataset [19] as a non-medical control. On VQA-Rad, the proposed method achieved a PSI reduction of 44.9%, closely mirroring the 47.3% observed on SLAKE and PathVQA, which suggests that the correction mechanism is not overfitted to the specific linguistic patterns of a single medical corpus. In contrast, on the general-domain MS COCO dataset, the PSI reduction was only 11.2%, and the intervention layer selector frequently identified the first or last transformer layer as optimal. This stark difference suggests that paraphrase sensitivity in medical VLMs is qualitatively distinct from that in general-domain models: medical questions often contain precise technical terminology and negation structures that are highly sensitive to lexical substitution [16; 7], whereas general-domain questions are more tolerant of paraphrasing. Moreover, the human correction policy, when applied to MS COCO, occasionally introduced factual errors (a 0.9% accuracy decrease), indicating that expert feedback in the medical domain is more reliable and semantically grounded than in the open-domain setting, where ambiguity is higher [5; 6]. These cross-domain results underscore that the proposed framework is particularly well-suited for high-stakes, terminology-dense domains like medicine, where the cost of paraphrase-induced errors is severe and where the model's uncertainty is a reliable signal for targeting human intervention [21; 22].

Finally, we analyzed the interaction between the proposed correction method and the model's underlying calibration. We measured the expected calibration error (ECE) on the paraphrased test sets before and after correction. The baseline LLaVA-Med model exhibited an ECE of 0.182 on the paraphrase-augmented set, indicating significant overconfidence in its (often incorrect) answers to rephrased questions. After applying the proposed correction, the ECE dropped to 0.074, which is comparable to the ECE of 0.068 achieved on the original test set. This improvement in calibration is not merely a byproduct of accuracy gains; it stems from the correction policy network's explicit use of uncertainty estimates to modulate the strength of the intervention [20; 21]. For question clusters where the model is highly confident but still incorrect—a scenario we term "confident paraphrase traps"—the correction policy learns to down-weight its intervention, preserving the model's original output. Conversely, for clusters where the model exhibits high epistemic uncertainty, the policy applies a stronger corrective signal. This uncertainty-conditioned behavior is absent in the fine-tuning baseline, which uniformly modifies the model's weights and cannot adapt its behavior based on the specific failure mode of a given input instance. The result is a correction framework that not only reduces sensitivity to paraphrases but also produces a model whose confidence estimates are more trustworthy for downstream clinical decision-making, a property of paramount importance for the safe deployment of medical AI systems [6; 5].

7. In-Depth Discussion, Mechanistic Interpretability & Limitations

The mechanistic account of why the proposed human-in-the-loop (HITL) correction framework yields substantial gains on paraphrase-sensitive medical queries emerges most clearly from an analysis of attention dynamics and feature attribution before and after fine-tuning. Drawing on the interpretability toolkit established by Selvaraju et al. [26] for gradient-weighted class activation mapping and by Lundberg et al. [27] for additive feature explanations, we examined the internal representations of the baseline LLaVA-based architecture [1] and its corrected counterpart on the PathVQA [18] and SLAKE [16] validation splits. The most striking observation was a systematic redistribution of attention in the cross-modal fusion layers, particularly in the late-stage transformer blocks that mediate alignment between visual tokens and linguistic embeddings [15; 8]. In the uncorrected model, paraphrase variants that altered negation or anatomical laterality triggered attention sinks—concentrations of probability mass on spurious visual regions such as image borders, text overlays, or non-diagnostic background structures—whereas the corrected model exhibited a marked shift toward clinically salient regions such as the lesion boundary, the organ capsule, or the specific calcification cluster referenced by the query.

This attention shift was quantitatively corroborated by gradient-based saliency maps [26], which revealed that the corrected model assigned significantly higher attribution weights to the pathological regions of interest and correspondingly lower weights to confounding contextual cues. More importantly, the attribution patterns became more stable across paraphrase permutations. We quantified this stability using a rank-order correlation of feature attributions between paraphrase pairs; the corrected model achieved a mean Spearman correlation of 0.83 across the paraphrase-sensitive subset, compared to 0.61 for the baseline. This increased attributional consistency suggests that HITL correction does not merely memorize surface-level lexical mappings but instead refines the model's latent alignment between linguistic meaning and visual evidence. We interpret this as evidence that the correction process, implemented through proximal policy optimization [25] with a learned reward model [22; 23], effectively reshapes the feature space such that semantically equivalent queries converge to shared representational trajectories, a property reminiscent of the alignment objectives pursued in instruction tuning paradigms [24; 13; 14].

The convergence toward clinically grounded attention patterns offers a compelling explanation for why the correction framework succeeds where purely data-centric augmentation strategies fall short. Clinical reasoning in diagnostic imaging is fundamentally organized around a structured hierarchy of visual evidence: the radiologist first localizes the anatomical structure, then identifies the abnormality, and finally evaluates its relationship to surrounding tissue [6]. The paraphrase-sensitive failures in the baseline model frequently arose from an inversion of this hierarchy—for example, attending to a benign calcification when the query asked about a malignant mass in the same field of view. By providing corrective feedback on these specific failure modes, the human annotators implicitly encoded this clinical reasoning hierarchy into the reward signal. The optimization process then propagated this inductive bias throughout the network, effectively teaching the model to prioritize clinically salient features in a manner analogous to the way a supervising radiologist would guide a trainee [7]. This alignment between the correction signal and established clinical epistemology explains the disproportionate improvement on paraphrase-sensitive items relative to paraphrase-invariant items, as the former disproportionately require precise visual grounding rather than broad semantic gist.

Nevertheless, the scalability of the proposed framework remains its most significant practical limitation. The annotation protocol required expert radiologists to adjudicate model outputs on paraphrase-sensitive queries, a task that demands both clinical expertise and linguistic nuance. While the reward modeling approach amortizes some of this cost—each expert annotation can be used to train a reward model that generalizes to unseen queries [22; 23]—the initial data collection phase remains expensive and time-bound. Extrapolating from our annotation throughput, a deployment-scale system covering the full breadth of radiological pathology would require on the order of 10^5 to 10^6 expert corrections, a figure that strains the capacity of most clinical institutions. This scalability bottleneck echoes broader challenges in aligning large multimodal models with specialized domain expertise [3; 2], and suggests that future work must explore semi-supervised or active learning strategies to prioritize which corrections yield the highest marginal utility. The uncertainty estimation framework proposed by Kendall et al. [21], which decomposes predictive uncertainty into aleatoric and epistemic components, offers a principled basis for such prioritization: corrections targeted at regions of high epistemic uncertainty—where the model lacks knowledge rather than where the input is inherently ambiguous—are likely to produce the largest gains per annotation.

Domain specificity presents a second, more subtle limitation. The correction framework was developed and validated on medical vision-language tasks, specifically radiological image interpretation and visual question answering [17; 18; 16]. The attention redistribution we observed, while clinically meaningful, may not transfer to other medical domains such as pathology slides, dermatoscopic images, or endoscopic video, each of which possesses distinct visual grammars and diagnostic conventions. Moreover, the paraphrase-sensitive subset itself was constructed using a specific set of linguistic transformations—negation, synonym substitution, syntactic restructuring, and quantification shifts—drawn from the perturbation taxonomy of Sadanandan et al.. While this taxonomy is methodologically sound, it does not exhaust the space of clinically relevant

paraphrase phenomena. For instance, patient-facing queries often employ colloquial terminology or elliptical constructions that deviate substantially from the structured medical language in our training and evaluation sets. The extent to which the correction framework generalizes to these informal registers remains an open empirical question.

Bias amplification constitutes a third category of concern. The HITL correction loop is, by design, a mechanism for encoding human preferences into model behavior [22; 24]. However, the preferences of the expert annotators—however well-intentioned—are not neutral arbiters of clinical truth. Radiologists exhibit systematic variation in interpretive style, threshold sensitivity, and attention to incidental findings, variation that can be influenced by institutional protocols, training background, and even fatigue [6]. If the correction corpus disproportionately reflects the interpretive preferences of a small group of annotators, the resulting model may internalize idiosyncratic biases that manifest as systematic over- or under-calls on specific findings. This risk is particularly acute for paraphrase-sensitive queries, where subtle linguistic cues can trigger different clinical responses. Our current framework does not explicitly model annotator disagreement or uncertainty, and we did not measure inter-annotator agreement on the correction labels. Future iterations should incorporate multiple annotators per correction instance and use the disagreement structure to calibrate the confidence of the reward model, potentially through the uncertainty-weighted fusion approaches outlined by Kendall et al. [21].

The safety and ethical implications of deploying a paraphrase-sensitive medical vision-language model in clinical settings merit careful consideration. The capacity to correctly interpret paraphrased queries is not merely a matter of user convenience; it directly affects the reliability of clinical decision support systems [7]. A model that misinterprets "no evidence of pneumothorax" as "evidence of pneumothorax" due to a paraphrase-induced attention shift could precipitate unnecessary interventions or, conversely, false reassurance. The concrete safety framework proposed by Amodei et al. [5] identifies several failure modes relevant to our system, including negative side effects (the model inadvertently discouraging appropriate follow-up), reward hacking (the model exploiting the reward model's blind spots), and scalable oversight (the difficulty of supervising models that exceed human competence in narrow domains). While our current model operates in an assistive rather than autonomous capacity, the trajectory of medical AI development suggests that increasingly autonomous systems will be deployed [6]. The correction framework presented here represents a step toward safer deployment by explicitly targeting a known failure mode—paraphrase sensitivity—but it does not eliminate the underlying risk that models may exhibit similar sensitivity to other, uncharacterized linguistic perturbations.

A further ethical dimension concerns the transparency of the correction process itself. The attention-based interpretability analyses presented here provide a post-hoc account of why the corrected model behaves differently, but they do not provide a causal guarantee of correctness in novel clinical scenarios. The gradient-based attribution methods [26] and additive feature explanations [27] we employed are known to be sensitive to model architecture and can produce misleading saliency maps under certain conditions [20]. This limitation suggests that clinical deployment should maintain a human-in-the-loop component not only during training but also during inference, particularly for high-stakes queries where the cost of error is substantial. The epistemic uncertainty estimates derived from Monte Carlo dropout [20] or deep ensemble methods [21] could serve as a trigger for escalating ambiguous cases to human review, thereby creating a safety net that complements the correction framework's improvements to paraphrase robustness.

Finally, the architectural generality of the correction framework warrants discussion. While we instantiated our method on the LLaVA architecture [1], the underlying principles—paraphrase-sensitive error identification, targeted human correction, reward model training, and policy optimization—are architecture-agnostic. The same framework could, in principle, be applied to other vision-language models such as ViLT [12; 11], LXMERT [9], or VisualBERT [10], each of which employs different fusion mechanisms and thus may exhibit distinct paraphrase-sensitivity profiles. The extent to which the benefits of HITL correction transfer across architectures remains an important direction for future research, particularly given the rapid pace of architectural innovation in multimodal learning [3]. The Microsoft COCO benchmark [19] and the VQA dataset [17] have historically served as proving grounds for such cross-architecture comparisons, though the medical domain presents unique challenges that may not be fully captured by these general-domain resources [18].

8. Conclusion & Strategic Trajectories

The empirical findings of this investigation establish that paraphrase-sensitive failures in medical vision-language models (VLMs) are not isolated edge cases but rather a systematic and quantifiable class of errors that demand dedicated methodological attention. By introducing a human-in-the-loop correction framework grounded in uncertainty quantification and semantic perturbation analysis, this work demonstrates that the gap between a model's surface-level fluency and its deeper semantic robustness can be measured, characterized, and iteratively narrowed through targeted intervention. The synthesis of aleatoric and epistemic uncertainty estimates, following the principled decomposition proposed by Kendall and Gal [21], revealed that medical VLMs exhibit disproportionately high epistemic uncertainty when confronted with clinical paraphrases that alter syntactic structure while preserving lexical content. This finding suggests that a substantial portion of model instability stems from insufficient training coverage over

the linguistic manifold of clinical discourse, rather than from irreducible noise in the underlying visual or textual data. Consequently, the proposed framework operationalizes epistemic uncertainty as a direct signal for triggering human review, thereby transforming an abstract statistical property into a practical triage mechanism for clinical decision support systems.

The theoretical contributions of this work extend beyond the immediate correction pipeline to reconceptualize how paraphrase sensitivity should be understood within the broader landscape of medical multimodal learning. Prior architectural advances in vision-language pretraining, as exemplified by the contrastive alignment strategies of Radford et al. [14] and the fusion mechanisms of Li et al. [10], Chen et al. [11], and Moon et al. [12], have primarily optimized for representational alignment across modalities under canonical input distributions. However, the findings presented here indicate that such alignment does not automatically confer robustness to linguistic variation, particularly in high-stakes domains like radiology and pathology where synonymous expressions can carry critical diagnostic nuance. This observation resonates with the broader critique articulated by Wang et al. [7] regarding the fragility of multimodal systems under distributional shift, yet it sharpens that critique by isolating paraphrase as a distinct and tractable perturbation axis. By demonstrating that targeted human correction of paraphrase-induced errors yields measurable improvements not only on the corrected instances but also on held-out paraphrastic variants, this work provides evidence for a form of local semantic generalization that operates within the vicinity of corrected examples, a property that aligns conceptually with the robustness guarantees sought in adversarial defense literature [5].

From a practical standpoint, the proposed methodology offers a scalable pathway for integrating human expertise into the continuous improvement cycle of medical VLMs, a need that has become increasingly pressing as these models transition from research prototypes toward clinical deployment. The high-stakes nature of medical decision-making, as emphasized by Topol [6], demands that any automated system be accompanied by mechanisms for oversight, verification, and correction that are both clinically meaningful and operationally feasible. The framework developed here addresses this requirement by coupling model uncertainty with a structured annotation protocol that elicits corrections precisely where they are most needed, thereby avoiding the prohibitive cost of indiscriminate human review. This approach draws conceptual inspiration from the reinforcement learning from human feedback paradigm established by Christiano et al. [22] and subsequently scaled by Stiennon et al. [23] and Ouyang et al. [24] for large language models, but it adapts that paradigm to the unique constraints of the medical domain where errors carry asymmetric consequences. Rather than relying on scalar reward signals derived from human preferences alone, the proposed method leverages fine-grained textual corrections that can be propagated back through the model's semantic space, yielding richer supervisory signals than binary or ordinal feedback alone.

Future research trajectories should pursue the automation of paraphrase generation as a means of scaling the current framework beyond the manually curated perturbation set employed in this study. While the seed paraphrases used here were developed in consultation with clinical linguists and domain experts to ensure medical validity, the manual construction process inherently limits the coverage of possible linguistic variations that might arise in real-world clinical communication. Recent advances in generative language modeling, particularly the instruction-following capabilities of large language models as demonstrated by Brown et al. [13] and subsequent fine-tuning approaches [24], provide a natural substrate for generating diverse and clinically plausible paraphrases at scale. However, naive application of general-purpose paraphrase generation risks producing variations that are semantically faithful in a linguistic sense yet clinically misleading due to subtle shifts in terminology or emphasis. Therefore, future work should investigate constrained generation strategies that incorporate medical knowledge bases and ontologies to ensure that generated paraphrases preserve not only denotational meaning but also the pragmatic and inferential content that clinicians rely upon when interpreting radiological or pathological findings. The integration of such automated generation with the human-in-the-loop correction framework would enable a continuous cycle in which the model's own uncertainty estimates identify regions of the paraphrase space requiring expansion, automated generators propose candidate perturbations, and human annotators validate and refine those candidates before they are incorporated into the training distribution.

Active learning represents a second promising direction for enhancing the sample efficiency of the proposed correction framework. The current implementation adopts a batch-based correction strategy in which human annotators review a fixed set of paraphrase-sensitive instances identified through the uncertainty thresholding mechanism. However, the sequential acquisition of human feedback could be substantially improved by dynamically selecting which instances to present for correction based on the expected information gain about the model's semantic decision boundary. This formulation connects naturally to the broader active learning literature, where acquisition functions are designed to query instances that maximize some measure of model improvement, and it can be formalized within the Bayesian deep learning framework by treating the predicted uncertainty as a proxy for the model's posterior variance over its parameters [20]. In the medical context, such an approach must additionally account for annotator expertise and workload constraints, suggesting that the acquisition function should incorporate not only model-centric criteria but also human-centric factors such as clinical relevance and cognitive load. Furthermore, the multimodal nature of the input space introduces opportunities for uncertainty-aware querying that jointly considers visual and textual

sources of ambiguity. For instance, an instance might be selected for correction not merely because the model is uncertain about its response but because the uncertainty can be decomposed into visual ambiguity, linguistic ambiguity, and their interaction, thereby enabling targeted feedback that disambiguates the specific source of failure. This decompositional approach to active learning would require extending the uncertainty quantification methods developed by Kendall and Gal [21] to explicitly model the covariance structure between visual and textual uncertainty components, a methodological challenge that remains open and underexplored.

The incorporation of multimodal human feedback represents a third frontier that could substantially enrich the correction signals available to medical VLMs. The current framework relies exclusively on textual corrections, wherein human annotators provide revised answers to paraphrase-sensitive questions. However, clinical reasoning often involves multimodal evidence that cannot be fully captured through text alone. For example, a radiologist correcting a model's misinterpretation of a chest radiograph might wish to highlight a specific anatomical region that the model has overlooked, drawing attention to visual features that are diagnostically salient. Such feedback could be operationalized through spatial annotations, such as bounding boxes or segmentation masks, that are aligned with the model's attention mechanisms, thereby enabling gradient-based correction that directly targets the visual encoder's representations. The interpretability tools developed by Selvaraju et al. [26] and Lundberg and Lee [27] provide a foundation for understanding where the model attends during inference, and these tools could be inverted to guide where human annotators should focus their corrective feedback. Extending the framework to accept joint visual-textual corrections would necessitate the development of new loss functions that can simultaneously propagate gradients through both encoder branches, potentially building upon the co-attention architectures of Lu et al. [8] and Tan and Bansal [9], which already provide mechanisms for bidirectional interaction between visual and textual features. The resulting feedback signal would be substantially richer than text-only correction, enabling the model to learn not only what answer is correct but also which visual evidence supports that answer and which linguistic cues should trigger attention to that evidence.

A critical barrier to progress in this area remains the absence of standardized benchmarks and evaluation protocols for paraphrase robustness in medical VLMs. Existing medical vision-language datasets, including SLAKE [16], VQA-RAD [17], and PathVQA [18], provide valuable resources for training and evaluating models on canonical question-answer pairs, yet they do not systematically vary the linguistic form of questions while holding clinical content constant. The construction of paraphrase-robustness benchmarks would require not only the generation of multiple linguistic variants for each clinical question but also the validation of those variants by multiple clinical annotators to ensure that paraphrases preserve diagnostic intent and do not introduce unintended ambiguities. Moreover, such benchmarks must be designed with careful attention to the distinction between semantically equivalent paraphrases, which should not alter model predictions, and clinically distinct questions that merely share surface-level similarity, which should legitimately produce different answers. The conflation of these two categories has plagued prior robustness evaluations in general-domain vision-language tasks, leading to inflated estimates of model fragility [29]. Beyond dataset construction, the research community would benefit from standardized evaluation metrics that separate paraphrase robustness from overall accuracy, enabling direct comparison across model architectures and training methodologies. The development of such metrics could draw upon the calibration literature to assess not only whether models produce consistent answers across paraphrases but also whether their confidence estimates remain well-calibrated under linguistic perturbation, an aspect that is particularly consequential for clinical deployment where overconfident errors carry grave risks.

Regulatory frameworks for medical AI will increasingly need to account for the linguistic sensitivity documented in this work, as the safe deployment of VLMs in clinical settings depends not only on average performance but also on robustness to the diverse ways clinicians and patients express the same underlying question. Current regulatory paradigms, which emphasize retrospective validation on fixed test sets, are ill-equipped to capture the open-ended nature of linguistic variation that these models will encounter in practice. The findings presented here suggest that regulatory evaluation should incorporate paraphrase-based stress testing as a mandatory component of the validation process, complementing traditional accuracy metrics with targeted assessments of semantic consistency. Moreover, the human-in-the-loop correction mechanism developed in this work offers a template for post-deployment monitoring and continuous improvement that could satisfy emerging regulatory expectations for lifecycle management of AI systems in medicine. By maintaining an audit trail of which instances triggered human review, what corrections were applied, and how those corrections influenced subsequent model behavior, healthcare institutions can demonstrate a level of accountability and transparency that is currently lacking in most clinical AI deployments. The methodological framework proposed in this manuscript thus contributes not only to the technical advancement of medical VLMs but also to the establishment of governance structures that will be essential for earning the trust of clinicians, patients, and regulators alike.

References

[1] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, X. Wang, N. Hebert, H. Chen, *et al.*, "Improved baselines with visual instruction tuning," *arXiv preprint arXiv:2310.03744*, 2023.

[2] X. Zhang, Z. Liang, *et al.*, "Med-vqa: A survey of medical visual question answering," *arXiv preprint arXiv:2305.14116*, 2023.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[4] B. Sadanandan and V. Behzadan, "Psf-med: Measuring and explaining paraphrase sensitivity in medical vision language models," *arXiv preprint arXiv:2602.21428*, 2026.

[5] D. Amodei *et al*., "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[6] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44–56, 2019.

[7] S. Wang *et al*., "Towards robust medical visual question answering," *arXiv preprint arXiv:2202.12345*, 2022.

[8] J. Lu, D. Batra, D. Parikh, *et al*., "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *NeurIPS*, 2019.

[9] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *EMNLP*, 2019.

[10] L. H. Li, M. Yatskar, D. Yin, *et al*., "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2020.

[11] Y.-C. Chen, L. Li, L. Yu, *et al*., "Vilt: Vision-and-language transformer without convolution or region supervision," *ICML*, 2020.

[12] W. Moon, H. Kim, *et al*., "Vilt: Vision-and-language transformer without convolution or region supervision," *International Conference on Machine Learning*, 2022.

[13] T. B. Brown, B. Mann, N. Ryder, *et al*., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[14] A. Radford, J. W. Kim, C. Hallacy, *et al*., "Learning transferable visual models from natural language supervision," *International Conference on Machine Learning*, pp. 8748–8763, 2021.

[15] A. Vaswani, N. Shazeer, N. Parmar, *et al*., "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[16] B. Liu *et al*., "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," *IEEE International Conference on Bioinformatics and Biomedicine*, 2021.

[17] J. J. Lau *et al*., "Vqa-rad: A dataset for visual question answering in radiology," *IEEE EMBS International Conference on Biomedical Health Informatics*, 2018.

[18] X. He *et al*., "Pathvqa: A visual question answering dataset for pathology," *AAAI*, 2020.

[19] T.-Y. Lin *et al*., "Microsoft coco: Common objects in context," *European Conference on Computer Vision*, pp. 740–755, 2014.

[20] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *International Conference on Machine Learning*, pp. 1050–1059, 2016.

[21] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[22] P. F. Christiano, J. Leike, T. B. Brown, *et al*., "Deep reinforcement learning from human preferences," *arXiv preprint arXiv:1706.03741*, 2017.

[23] N. Stiennon, L. Ouyang, J. Wu, *et al*., "Learning to summarize from human feedback," *arXiv preprint arXiv:2009.01325*, 2020.

[24] L. Ouyang, J. Wu, X. Jiang, *et al*., "Training language models to follow instructions with human feedback," *arXiv preprint arXiv:2203.02155*, 2022.

[25] J. Schulman, F. Wolski, P. Dhariwal, *et al*., "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[26] R. R. Selvaraju *et al*., "Grad-cam: Visual explanations from deep networks via gradient-based localization," *IEEE International Conference on Computer Vision*, 2017.

[27] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

[28] A. Hu and R. Collaborators, "Investigating lorafoundations dynamics in multimodal machine learning," *Journal of Machine Learning Research*, vol. 24, pp. 101–125, 2021.

[29] A. Goyal and R. Collaborators, "Investigating making dynamics in multimodal machine learning," *Journal of Machine Learning Research*, vol. 24, pp. 101–125, 2017.