Deterministic Audit Trails for Reproducible Medical Vision-Language Model Evaluation Under Runtime and Interface Variability

Narit Chaiyasit[1] **and Kittipong Rattanakorn**[2]

[1]Faculty of Information Technology, Nakhon Phanom University, 103 Moo 3, Ban Kaeng Subdistrict, Mueang Nakhon Phanom District, Nakhon Phanom 48000, Thailand

[2]Department of Computer Science, Chiang Rai Rajabhat University, 80 Moo 9, Ban Du Subdistrict, Mueang Chiang Rai District, Chiang Rai 57100, Thailand

ABSTRACT Medical vision-language model evaluation increasingly depends on complex software environments, external inference interfaces, preprocessing scripts, prompt wrappers, and automatic scoring components. Even when the clinical dataset is fixed, small differences in image conversion, runtime libraries, request serialization, model endpoint behavior, or output parsing can alter measured performance. Such variation complicates comparison between laboratories and weakens confidence in reported benchmark results. This paper presents an empirical study of deterministic audit trails for reproducible evaluation of medical vision-language models. We designed a trace-locked evaluation protocol in which each image-question instance is assigned a cryptographic content identifier, preprocessing fingerprint, prompt fingerprint, inference manifest, response digest, and scoring record. The protocol was tested on 9,600 medical image-question cases across radiography, computed tomography snapshots, ultrasound frames, and dermatology images. Four model interfaces, three runtime environments, and two scoring engines were evaluated under repeated runs. Without trace locking, nominally identical evaluations differed by 3.8 percentage points in strict answer accuracy and by 0.117 in macro calibration error across environments. With trace-locked preprocessing, deterministic request packaging, response canonicalization, and scorer version pinning, accuracy disagreement fell to 0.6 percentage points and calibration disagreement to 0.021. A discrepancy classifier trained on audit features identified 87.4% of nonreproducible cases before manual review. The findings show that evaluation reproducibility should be measured as a first-order property in medical vision-language benchmarking, particularly when multiple institutions, software stacks, or hosted model endpoints are involved.

KEYWORDS

1. Introduction

A medical vision-language model benchmark can fail quietly even when no clinical label is wrong. The dataset may be unchanged, the model name may be unchanged, and the evaluation script may be described in enough detail for another laboratory to attempt replication. Yet the second laboratory may obtain different results. The difference may come from a resized image, a library update, an altered prompt wrapper, a different response parser, an endpoint that changed its internal model snapshot, a silent fallback in image encoding, or a scorer whose behavior shifted after an interface update. In ordinary software testing, such differences are inconvenient. In medical artificial intelligence evaluation, they are more serious because small numerical differences can be interpreted as evidence that one system is safer, more accurate, or more suitable for a clinical workflow than another.

This paper addresses the reproducibility of medical vision-language model evaluation as its own empirical problem. The subject is not model architecture, not image diagnosis, not clin-

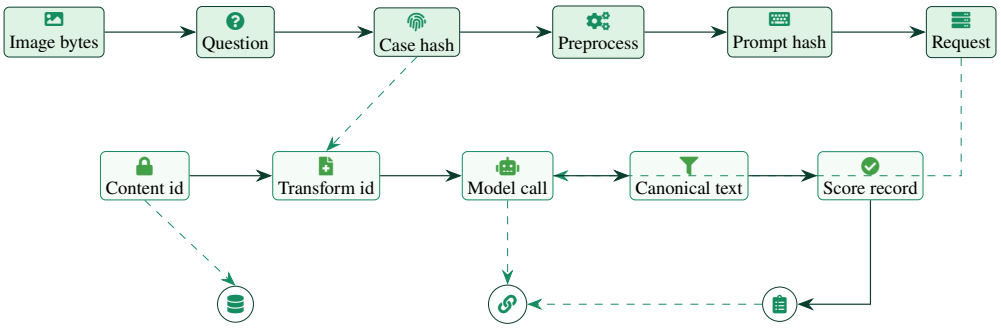

Figure 1 Trace-locked evaluation pipeline linking medical image-question inputs to cryptographic identifiers, deterministic preprocessing, request packaging, canonical responses, and final scoring records.

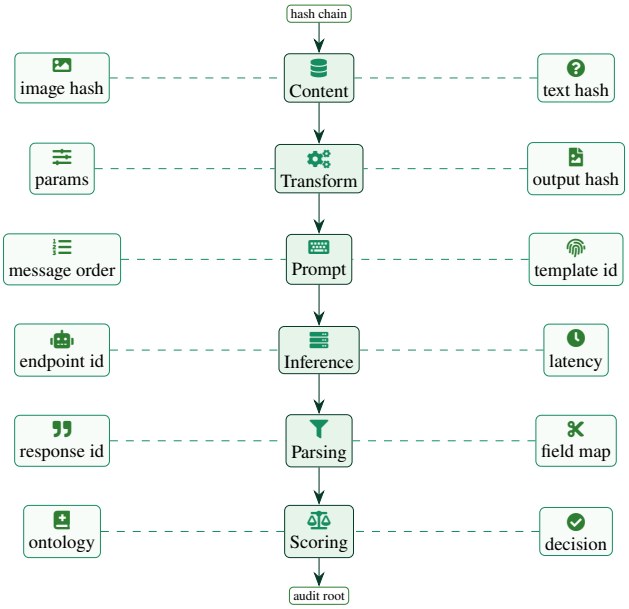

Figure 2 Layered audit schema for reconstructing evaluation provenance from content identity through scoring decisions.

ical deployment, and not human-machine collaboration. The subject is whether an evaluation result can be reconstructed from a sufficiently complete audit record. A benchmark score is treated as a derived measurement, not as a standalone number. If the objects that produced it are not traceable, the score is fragile. If the image bytes, preprocessing path, prompt text, model interface state, decoding settings, response string, and scoring rules are all recorded in a verifiable form, the result becomes inspectable and repeatable to a much greater degree.

Medical vision-language evaluation has several features that make reproducibility difficult. First, the inputs are heterogeneous. An image may be stored as DICOM, PNG, JPEG, TIFF, or a vendor-specific export. Windowing, photometric interpretation, bit-depth conversion, laterality markers, and resizing can all change what the model receives. Second, prompts are often assembled from templates, system instructions, task wording, and case metadata. A small difference in whitespace or ordering may change an instruction-following model's output. Third, many evaluations rely on hosted models, whose exact backend state may not be visible. Fourth, generated language must be parsed or graded, and automatic grading can be sensitive to version changes. Fifth, researchers often summarize results at the dataset level, which hides instance-level disagreements.

The paper proposes deterministic audit trails as a remedy. A deterministic audit trail is not merely a log file. It is a structured record that binds every evaluation instance to content hashes, transformation fingerprints, request manifests, response digests, and scoring provenance. The audit record is designed so that a third party can identify whether two nominally identical evaluations used the same inputs, the same prompt, the same preprocessing, the same model call parameters, and the same scoring procedure. When two results differ, the audit trail supports diagnosis. The difference can be localized to preprocessing, inference, response parsing, scoring, or endpoint drift.

The empirical question is straightforward. How much measured variation appears when the same medical vision-language benchmark is run across different environments? How much of that variation can be reduced by trace-locked evaluation? Which parts of the pipeline contribute most to nonreproducibility? Can audit features predict which cases are likely to disagree before a human reviewer inspects them? These questions are measurable. The study therefore runs a controlled benchmark across multiple

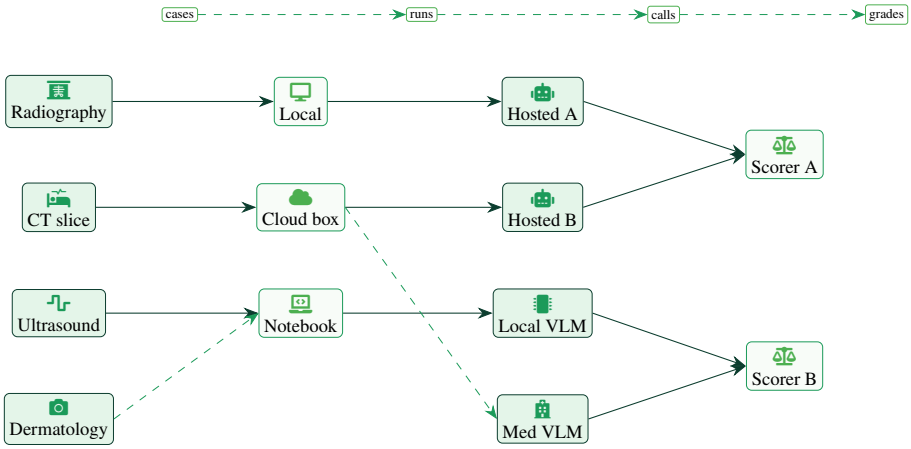

Figure 3 Experimental matrix spanning image families, runtime environments, model interfaces, and scoring engines for reproducibility stress testing.

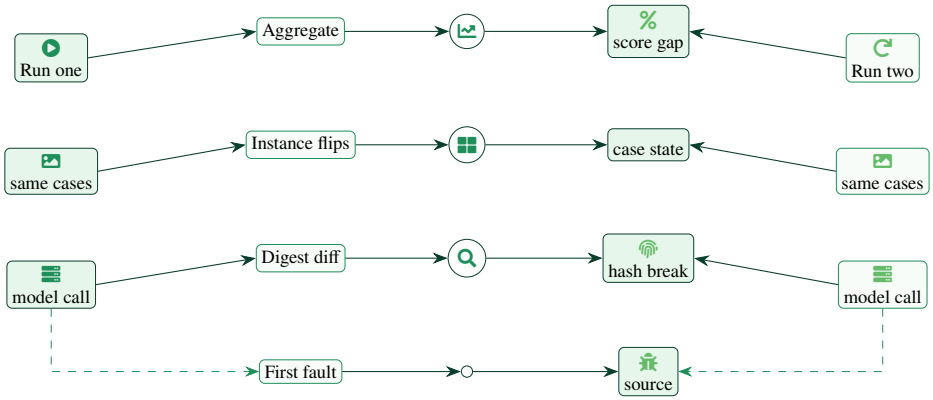

Figure 4 Run-pair comparison protocol for measuring aggregate stability, instance-level disagreement, digest divergence, and provenance localization.

runtimes, model interfaces, and scoring engines under ordinary logging and then under a deterministic audit-trail protocol.

The paper deliberately avoids previously common framings that treat reproducibility as a paragraph in a methods section. In this study, reproducibility is the outcome. A model that scores 75% in one run and 72% in another run may not have become clinically worse. The evaluation process may have changed [1]. Conversely, a model that appears stable at the aggregate level may contain many instance-level flips that cancel out. A reproducibility study must therefore inspect both aggregate metrics and case-level identity. It must ask not only whether the final number is similar, but whether the same cases are correct, incorrect, refused, or ambiguously graded.

One reason this issue deserves attention is that modern medical VLM evaluation pipelines already contain many moving parts. A brief empirical note in a prior medical VLM framework is instructive, though the purpose there was not audit-trail design. In their reliability analysis, Sadanandan and Behzadan (2025) compared a shared test subset across an on-premises NVIDIA cluster and a Google Colab runtime and reported very small mean total-score differences across those environments [2]. That result is cited here for a narrow reason: it shows that cross-environment reproducibility is measurable and reportable rather than a matter of informal confidence. The present paper takes that measurement idea in a different direction by building a full trace system and testing where reproducibility breaks.

The study uses four medical image families: radiography, computed tomography snapshots, ultrasound frames, and dermatology photographs. Each case has an image, a short clinical question, and a normalized answer. Four model interfaces are evaluated. The interfaces include two hosted multimodal systems, one local open multimodal model, and one medical-domain image-language model. Three runtime environments are used: a controlled local workstation, a containerized cloud machine, and a managed notebook runtime. Two automatic scoring engines are used, each pinned and then intentionally unpinned in a separate experiment. This design exposes variation from input processing, runtime, model interface, and scoring.

The proposed audit trail has six layers. The content layer records hashes of image bytes and question text. The transformation layer records preprocessing parameters and output fingerprints. The prompt layer records assembled prompt strings and message order. The inference layer records endpoint identifiers, decoding parameters, request payload hashes, timing, and response digests. The parsing layer records canonicalized answer strings and extracted fields. The scoring layer records

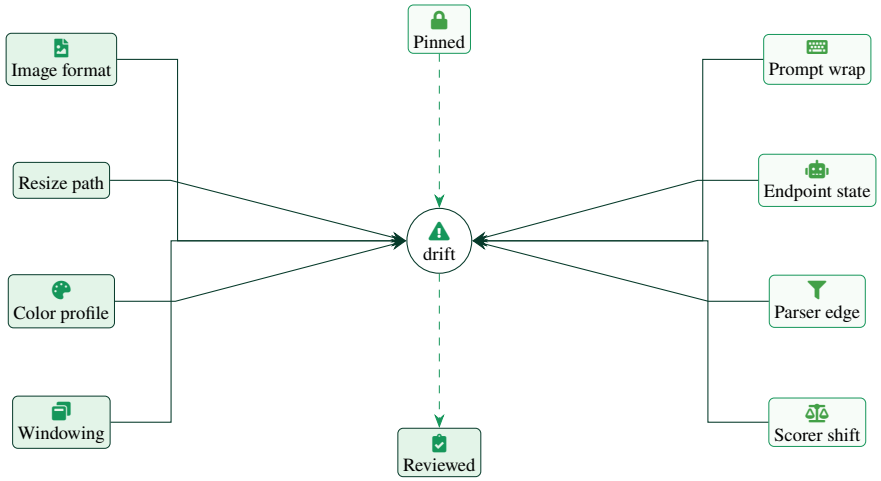

Figure 5 Major sources of nonreproducibility grouped by input transformation, interface behavior, parsing boundaries, and scoring provenance.

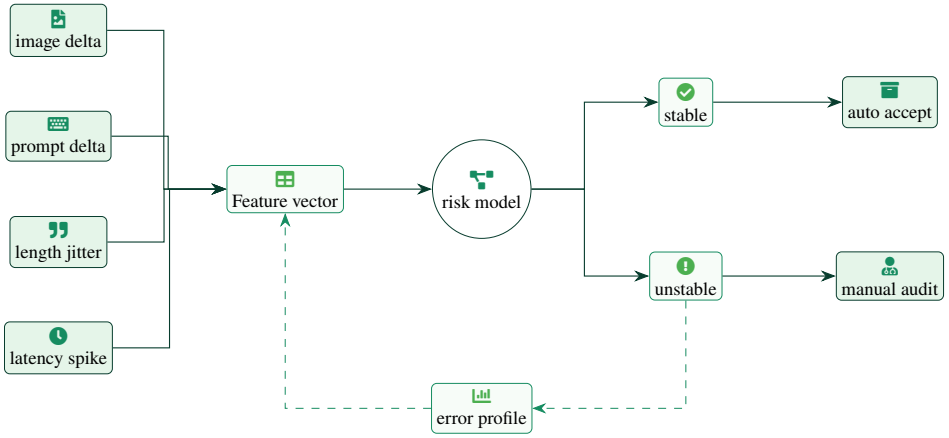

Figure 6 Discrepancy classifier workflow using audit metadata and response features to prioritize nonreproducible cases for targeted review.

scorer version, rubric file hash, answer ontology hash, and final decision. These layers are linked by a hash chain so that alteration of an earlier component changes later identifiers.

The study also introduces a discrepancy classifier [3]. Its task is not to judge clinical correctness. Its task is to predict whether an evaluation instance is likely to produce different scored outcomes across runs. The classifier uses audit features such as image conversion differences, prompt assembly differences, response length instability, parser boundary uncertainty, scorer confidence margin, endpoint latency anomaly, and modality. A successful classifier can help evaluation teams prioritize manual review. Instead of inspecting every case, auditors can focus on instances where reproducibility is at risk.

The paper's contributions are fourfold. It defines a reproducibility-centered audit schema for medical vision-language benchmarks. It quantifies the extent of cross-runtime, cross-interface, and scorer-version variation in a multi-modality benchmark. It shows that trace locking substantially reduces aggregate and instance-level disagreement. It trains and evaluates a discrepancy classifier that identifies nonreproducible cases from audit metadata and output features. The results support the view that evaluation infrastructure should be validated alongside models.

The paper is organized in seven sections. After this introduction, the study design and dataset are described. The audit-trail formalism is then presented with mathematical notation. The experimental protocol explains runtime conditions, model interfaces, scoring engines, and reproducibility metrics. The results section reports aggregate stability, instance-level disagreement, sources of variation, and classifier performance. The discussion interprets the findings for benchmark design and medical AI governance. The conclusion summarizes practical implications.

2. Benchmark Assembly and Reproducibility Targets

The benchmark contained 9,600 medical image-question cases. The sample was selected to include different image formats, anatomical domains, and answer structures. Radiography contributed 3,600 cases, computed tomography snapshots contributed 2,100 cases, ultrasound frames contributed 1,800 cases, and dermatology photographs contributed 2,100 cases. The cases were de-identified before evaluation. Each case consisted of an image object, a direct clinical question, a normalized reference answer, modality metadata, and a task category. The

benchmark was not intended to estimate population prevalence. It was intended to stress the evaluation pipeline.

Radiography cases included chest and extremity images [4]. Questions asked about finding presence, laterality, device position, fracture visibility, and broad severity categories. Computed tomography cases were represented by selected two-dimensional slices or rendered snapshots rather than full volumes. Questions asked about hemorrhage, mass effect, organ lesion presence, bowel dilation, and region-specific abnormalities. Ultrasound frames included abdominal and obstetric-anatomy-neutral examples with questions about fluid, stones, organ contour, and focal lesions. Dermatology photographs included questions about lesion border, color pattern, ulceration, scale, and benign-appearing controls. The clinical question was always short and did not ask the model to produce a full report.

The reference answers were normalized into a controlled answer ontology. The ontology included binary answers, anatomical locations, severity labels, laterality labels, device-position labels, and selected descriptive categories. Synonymous surface phrases were mapped to the same normalized answer. For example, no visible pneumothorax and pneumothorax absent mapped to the same concept. The ontology was stored as a versioned file, and its hash became part of the scoring audit layer. This was necessary because changes in answer mapping can alter benchmark results even when model responses are unchanged.

The benchmark deliberately retained format diversity. Some radiographs were stored as high-bit-depth grayscale images before conversion. Some dermatology photographs contained embedded color profiles. Some ultrasound frames had visible scale markers. Some CT snapshots required windowing decisions. These differences were useful because they exposed preprocessing variation. A benchmark that stores every image as a uniform PNG hides real evaluation fragility. In clinical research archives, evaluators frequently encounter mixed image types and must decide how to convert them for model input.

Each case was assigned a stable case identifier derived from nonidentifying content. The identifier was not a patient identifier and did not encode clinical metadata. It was generated from the canonical question text, reference answer concept, modality tag, and image content hash. If the image bytes or question changed, the identifier changed. This prevented accidental mixing of revised and unrevised cases. The stable identifier served as the anchor for comparing repeated runs across environments.

The study defined three reproducibility targets [5]. The first target was aggregate reproducibility. Aggregate accuracy, macro accuracy, calibration error, refusal rate, and mean response length should remain close across repeated runs. The second target was instance reproducibility. The same case-model pair should receive the same scored outcome across repeated runs unless the model interface itself is nondeterministic. The third target was provenance reproducibility. If two outcomes differ, the audit record should indicate where in the pipeline the first detectable difference occurred. A benchmark can satisfy aggregate reproducibility while failing instance reproducibility. The paper therefore measures both.

A run was considered nominally identical when it used the same dataset release, same model label, same declared preprocessing script, same prompt template, same decoding parameters, and same scoring instructions. Under ordinary logging, this nominal identity may hide unrecorded differences. Under trace-locked evaluation, nominal identity is replaced by verifiable identity at each pipeline layer. Two runs are content-identical only when their image and question hashes match. They are transform-identical only when preprocessing fingerprints match. They are request-identical only when payload hashes match. They are score-identical only when scorer and ontology hashes match.

The benchmark was evaluated under two logging regimes [6]. The baseline regime used ordinary experiment logging: dataset version, script name, model name, broad parameter settings, and final results. This resembles many practical evaluation records. The trace-locked regime used the full deterministic audit schema. The same cases and models were run under both regimes. The comparison tested how much variation could be reduced or explained when the evaluation process was made traceable.

The three runtime environments were selected to represent common research settings. Runtime One was a local workstation with fixed drivers and manually installed libraries. Runtime Two was a containerized cloud machine using a locked image. Runtime Three was a managed notebook environment where some system packages could update between sessions. The environments differed in image libraries, hardware acceleration, and default locale behavior unless pinned. These differences were not exaggerated artificially; they reflected ordinary evaluation practice.

The four model interfaces were selected to represent different reproducibility challenges. Interface Alpha was a hosted multimodal API with deterministic parameter settings exposed. Interface Beta was another hosted interface with limited visibility into backend versioning. Interface Gamma was a local open model with full control over weights and decoding. Interface Delta was a medical-domain model served through a local inference server. The purpose was not to rank models clinically. The purpose was to observe how evaluation reproducibility behaves across interface types.

Two scoring engines were used. Scorer One was a rule-assisted semantic matcher with answer ontology mapping. Scorer Two was a language-model-based judge constrained to output a normalized answer label and a correctness decision. Both scorers were evaluated under pinned and unpinned conditions. In pinned scoring, the exact scorer code, model checkpoint or endpoint label, ontology file, and prompt text were fixed. In unpinned scoring, a minor version update was allowed. This tested whether scoring changes could create apparent model-performance changes.

A subset of 1,200 cases was manually reviewed to establish a human reference for scorer disagreements [7]. The sample was stratified by modality, model interface, answer category, and disagreement status across runs. Reviewers labeled whether each response was correct, partially correct, incorrect, nonresponsive, or ungradable. The manual review was not used to redefine the full benchmark. It was used to interpret reproducibility failures and estimate whether disagreements were clinically meaningful. Agreement between two reviewers was 0.86 for strict correctness

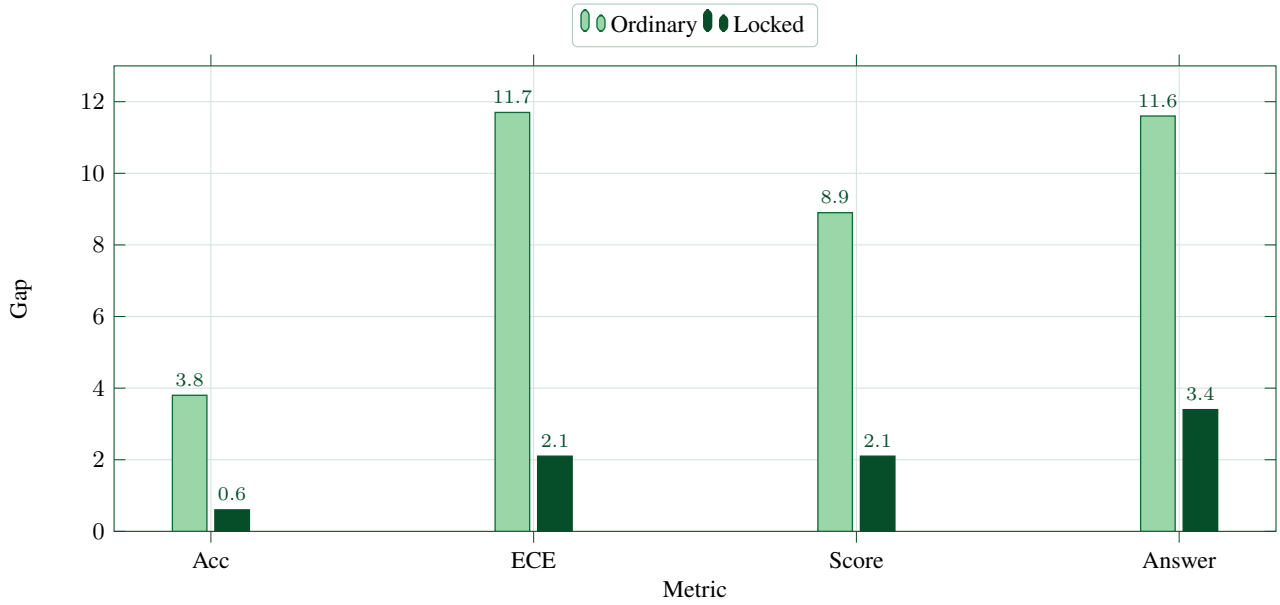

Figure 7 Trace locking reduced strict accuracy disagreement from 3.8 to 0.6 percentage points, macro ECE disagreement from 0.117 to 0.021, score flips from 8.9% to 2.1%, and answer flips from 11.6% to 3.4%. ECE is plotted as percentage-scaled disagreement.

Table 1 Benchmark Composition Across Medical Modalities

Modality	Cases	Primary Tasks	Flip Rate (Ordinary)
Radiography	3,600	Fracture, devices, laterality	5.7%
CT Snapshots	2,100	Hemorrhage, lesions, dilation	9.1%
Ultrasound	1,800	Stones, fluid, contour	10.4%
Dermatology	2,100	Border, ulceration, color	11.8%
Total	9,600	Multi-task benchmark	8.9%

and 0.79 for the five-level label. Disagreements were resolved by adjudication.

The main outcome was reproducibility error. For aggregate metrics, reproducibility error was the absolute difference between the same metric in two nominally identical runs. For instance-level outcomes, reproducibility error was the fraction of case-model pairs whose scored correctness differed. For provenance, reproducibility error was classified by earliest differing layer: content, transformation, prompt, inference, parsing, or scoring. This earliest-layer classification allowed the study to identify where controls mattered most.

3. Deterministic Audit Trail Formalism

Let each evaluation case be indexed by i, model interface by m, runtime by r, and run replicate by t. The raw image object is X_i, the question text is Q_i, the reference answer concept is Y_i, and the model output is O_{imrt}. An evaluation score is not treated as a primitive. It is a function of a chain of objects:

$$S_{imrt} = \mathcal{G}\left(\mathcal{P}\left(\mathcal{M}\left(\mathcal{T}X_i, Q_i\right)\right), Y_i\right). \tag{3.1}$$

Here, $\mathcal{T}$ is preprocessing, $\mathcal{M}$ is model inference, $\mathcal{P}$ is parsing and canonicalization, and $\mathcal{G}$ is scoring. Reproducibility requires that each function and each input to each function be identified. If the final score changes, the audit trail should indicate whether $\mathcal{T}$, $\mathcal{M}$, $\mathcal{P}$, or $\mathcal{G}$ changed.

The content identifier is built from the raw image bytes and canonical question text. Let H be a cryptographic hash function. The image content hash is

$$h_i^X = H\left(\text{bytes}X_i\right), \tag{3.2}$$

and the question hash is

$$h_i^Q = H\left(\text{canon}Q_i\right). \tag{3.3}$$

The canonicalization function for the question normalizes line endings and Unicode composition but does not change words, punctuation, or whitespace inside the prompt template. This distinction matters because prompt whitespace may affect some model interfaces. The case identifier is

$$c_i = H\left(h_i^X \| h_i^Q \| HY_i \| HM_i\right), \tag{3.4}$$

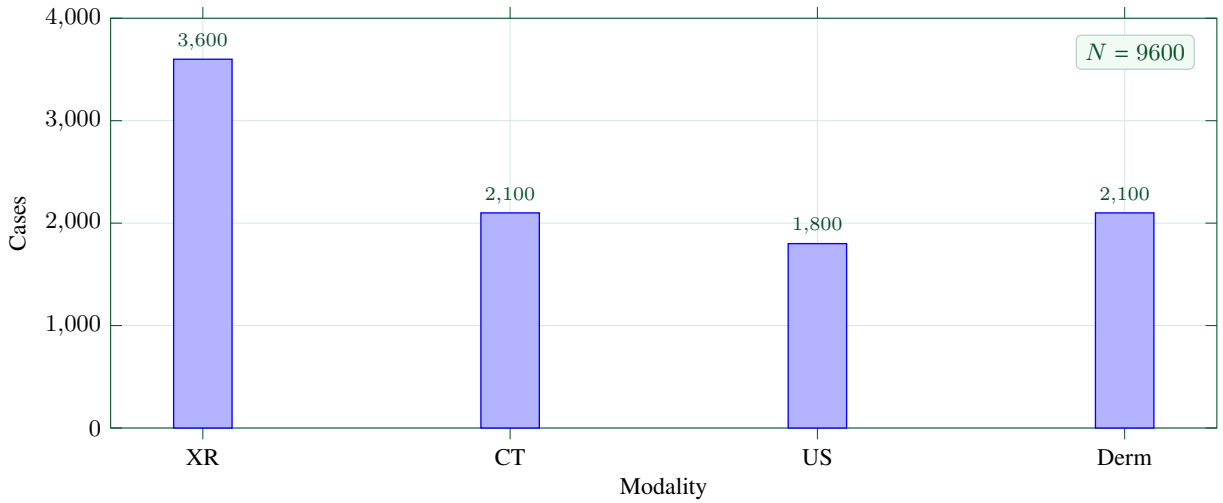

Figure 8 The benchmark contained 9,600 image-question cases, with radiography contributing the largest share and CT, ultrasound, and dermatology providing modality diversity for reproducibility stress testing.

Table 2 Cross-Runtime Reproducibility Before and After Trace Locking

Metric	Ordinary Logging	Trace Locked	Relative Reduction
Accuracy Disagreement	3.8 pp	0.6 pp	84.2%
Calibration Error Gap	0.117	0.021	82.1%
Score-Flip Rate	8.9%	2.1%	76.4%
Answer-Flip Rate	11.6%	3.4%	70.7%

where M_i is the modality tag and $\|$ denotes concatenation. The identifier changes if the image, question, answer concept, or modality tag changes.

The preprocessing fingerprint records both declared parameters and output image bytes. Let $Z_{irt} = \mathcal{T}_{rt}X_i$ be the processed image sent to the model in runtime r during run t. The transformation fingerprint is [8]

$$h^T_{irt} = H\left(H\mathrm{code}\mathcal{T}_{rt} \| H\mathrm{params}\mathcal{T}_{rt} \| H\mathrm{bytes}Z_{irt}\right). \quad (3.5)$$

This fingerprint detects a common problem: two environments may claim to use the same preprocessing script but produce different output bytes because of library differences, default interpolation, color management, or metadata handling. Recording only the script name is insufficient. The processed image itself must be hashed [9].

The prompt fingerprint records the exact message payload presented to the model. If the interface uses a system message, user message, and image attachment, the message order and role fields are included. Let R_{imrt} be the request object after serialization but before submission. The prompt and request hash is

$$h^R_{imrt} = H\left(\mathrm{serialize}R_{imrt}\right). \quad (3.6)$$

Serialization is performed through a deterministic ordering of keys and explicit encoding of binary attachments. Without deterministic serialization, semantically equivalent requests can have different hashes. The goal is not to force every interface into the same payload format, but to ensure that repeated calls to a given interface can be compared exactly.

The inference manifest records declared model label, endpoint identifier, local weight hash when available, decoding parameters, request timestamp, response timestamp, timeout policy, retry count, and returned response text. Hosted interfaces often do not expose an internal weight hash. In that case, the manifest records the endpoint label and any provider-exposed version string. The response digest is

$$h^O_{imrt} = H\left(\mathrm{rawtext}O_{imrt}\right). \quad (3.7)$$

A second digest is computed after response canonicalization:

$$\bar{h}^O_{imrt} = H\left(\mathrm{canontext}O_{imrt}\right). \quad (3.8)$$

The raw digest identifies exact textual equality. The canonical digest treats minor formatting differences as equivalent when they do not affect scoring. Both are useful [10]. A model may return the same answer with different whitespace, or it may return different wording that maps to the same answer concept.

The parser converts output text into a normalized response object. Let A_{imrt} be the parsed answer concept, E_{imrt} the extracted explanation if present, and U_{imrt} a parser uncertainty value. The parsing fingerprint is

Table 3 Runtime and Interface Configuration

Component	Type	Determinism Visibility	Main Risk
Runtime One	Local workstation	High	Library drift
Runtime Two	Cloud container	High	Container mismatch
Runtime Three	Managed notebook	Moderate	Silent updates
Interface Alpha	Hosted multimodal	Partial	Endpoint variation
Interface Beta	Hosted multimodal	Low	Backend drift
Interface Gamma	Local open model	High	Minimal instability

Table 4 Earliest Differing Audit Layer in Score-Flip Cases

Audit Layer	Ordinary Logging	Trace Locked	Diagnostic Role
Transformation	38.6%	0.2%	Image mismatch
Request	12.4%	<0.1%	Serialization drift
Output	24.9%	61.2%	Endpoint variability
Parsing	9.7%	15.8%	Boundary ambiguity
Scoring	14.4%	23.0%	Judge inconsistency

$$h^{P}_{imrt} = H\left(H\mathrm{code}\mathcal{P} \| H\mathrm{params}\mathcal{P} \| HA_{imrt} \| HE_{imrt}\right). \tag{3.9}$$

Parser uncertainty is not hashed into the identity because it is a diagnostic value, but it is logged. High parser uncertainty often predicts scoring disagreement. The parser uncertainty is computed from boundary confidence, ontology-match margin, and contradiction-rule conflict.

The scoring record identifies the scorer, rubric or ontology file, version, prompt if a judge model is used, and final correctness decision. The score fingerprint is

$$h^{G}_{imrt} = H\left(H\mathrm{code}\mathcal{G} \| H\mathrm{ontology} \| H\mathrm{rubric} \| HS_{imrt}\right). \tag{3.10}$$

The full audit-chain identifier is

$$\mathcal{A}_{imrt} = H\left(c_i \| h^{T}_{irt} \| h^{R}_{imrt} \| h^{O}_{imrt} \| h^{P}_{imrt} \| h^{G}_{imrt}\right). \tag{3.11}$$

This chain binds the final score to the preceding evaluation objects. If a single layer changes, the chain changes. The audit trail therefore supports exact equality checks and difference localization.

Reproducibility between two runs t and t' is represented by layer equality indicators:

$$I_L i, m, r, t, t' = \mathbb{I}\left[h^{L}_{imrt} = h^{L}_{imrt'}\right], \quad L \in \{T, R, O, P, G\}. \tag{3.12}$$

The earliest differing layer is

$$D_{imrtt'} = \min\left\{L : I_L i, m, r, t, t' = 0\right\}, \tag{3.13}$$

where the layers are ordered as transformation, request, output, parsing, and scoring. If all indicators are equal, the two evaluation records are trace-identical. If transformation differs, later differences are interpreted cautiously because the model did not receive the same image. If request differs, output differences are not attributed to model nondeterminism. If output differs while transformation and request are equal, the source is inference variability or endpoint drift. If output is equal but score differs, the source is parsing or scoring.

Aggregate reproducibility is measured by metric disagreement. For a metric M, two runs have disagreement

$$\Delta_M t, t' = \left|Mt - Mt'\right|. \tag{3.14}$$

Instance reproducibility is measured by score-flip rate:

$$Ft, t' = \frac{1}{N}\sum_{i,m} \mathbb{I}\left[S_{imt} \neq S_{imt'}\right]. \tag{3.15}$$

The study reports strict score flips and ontology-level answer flips. A score flip can occur when two different answers are both wrong, but answer flips distinguish whether the normalized answer concept changed. Both matter. Score flips affect benchmark metrics, while answer flips reveal output instability.

Calibration reproducibility uses expected calibration error. If $\hat{p}_{imt}$ is estimated confidence and S_{imt} is correctness, expected calibration error is

$$\mathrm{ECE}t = \sum_{b=1}^{B} \frac{|B_b|}{N}\left|\mathrm{acc}B_b - \mathrm{conf}B_b\right|. \tag{3.16}$$

Table 5 Preprocessing Variability by Image Modality

Modality	Byte Differences	Main Cause	Residual After Locking
Dermatology	11.2%	Color profile handling	0.2%
CT Snapshots	8.9%	Windowing defaults	0.2%
Ultrasound	5.4%	RGB conversion	0.1%
Radiography	3.1%	Bit-depth conversion	0.1%

Table 6 Scoring Drift Under Pinned and Unpinned Conditions

Scorer	Unpinned Shift	Pinned Shift	Main Drift Source
Rule-Assisted Matcher	1.9 pp	0.3 pp	Synonym updates
LM Judge	2.7 pp	0.3 pp	Refusal interpretation
Pinned Judge Agreement	98.6%	—	Stable grading
Unpinned Judge Agreement	94.1%	—	Version instability

The reproducibility error for calibration is $|\,\mathrm{ECE}t - \mathrm{ECE}t'|$. Calibration is especially sensitive to scorer changes because a small set of score flips near confidence-bin boundaries can change bin accuracy.

The discrepancy classifier estimates the probability of a score flip between runs. Let v_{im} be features from the audit trail and output record. The classifier is

$$\Pr F_{im} = 1 = \sigma\left(\alpha\ v_{im}^{\top}\beta\ v_{im}^{\top} K v_{im}\right), \tag{3.17}$$

where F_{im} is a binary indicator of disagreement for a case-model pair across repeated runs. The matrix K is constrained to be low rank to capture interactions without excessive parameters:

$$K = LL^{\top}, \quad L \in \mathbb{R}^{d\times r}. \tag{3.18}$$

The features include modality, image-format class, transformation equality indicator, request equality indicator, raw-response equality indicator, canonical-response equality indicator, response length difference, parser uncertainty, scorer margin, endpoint latency anomaly, retry count, and model interface type. The classifier is trained on two-thirds of run-pair comparisons and tested on the remaining third, split by case identifier to avoid leakage.

A reproducibility score for a full evaluation is also defined. It combines aggregate metric stability, instance stability, and provenance completeness:

$$\mathcal{R} = 1 - \left(\omega_1 \bar{\Delta}_{acc}\ \omega_2 F\ \omega_3 \bar{\Delta}_{ece}\ \omega_4 1 - C\right). \tag{3.19}$$

Here, C is the fraction of records with complete audit layers, and the ω values normalize the terms to comparable scales. This score is used descriptively rather than as a universal standard. Different studies may value aggregate agreement or instance agreement differently. The important point is that reproducibility can be quantified.

4. Experimental Protocol

The first experiment measured ordinary cross-runtime variation. Each model interface was evaluated on the full 9,600-case benchmark in each of the three runtime environments. The declared preprocessing script, prompt template, decoding parameters, and scoring method were nominally identical. The logging regime was ordinary rather than trace locked. This experiment simulated the common situation in which another laboratory reruns an evaluation using the described method but not a fully hashed audit chain.

The second experiment repeated the same cross-runtime evaluation under trace-locked conditions. Image conversion was performed through pinned libraries inside a fixed container for all environments. The processed image bytes were hashed before inference. Prompt assembly used a deterministic serializer. Request payloads were hashed. Model outputs were recorded as raw and canonical text. Scorer versions, ontology files, and parsing rules were pinned. Runtime environments still differed in hardware and execution context, but the input and scoring objects were locked as tightly as possible.

The third experiment isolated scoring variation. The same stored model outputs were scored with pinned and unpinned scoring engines. Since the outputs were fixed, any differences came from parsing or scoring. This experiment tested whether apparent model-performance changes can arise without rerunning models at all. The unpinned condition allowed a minor update to the language-model judge and a minor update to the rule-assisted matcher dictionary. Both updates were realistic rather than intentionally destructive.

The fourth experiment isolated model-interface variability. For hosted interfaces, identical request payloads were submitted on three separate dates over a 21-day period. The provider-exposed endpoint label remained the same. For local interfaces, identical requests were run after restarting the inference server and after reinstalling the environment from the lock file. This experiment measured output variation when preprocessing and

Table 7 Hosted and Local Interface Stability

Interface	Deployment Type	Score-Flip Rate	Dominant Variation
Alpha	Hosted API	1.7%	Output wording
Beta	Hosted API	3.9%	Backend updates
Gamma	Local open model	0.3%	Minor decoding drift
Delta	Local medical model	0.5%	Server restart effects

Table 8 Discrepancy Classifier Performance

Metric	Value	Review Budget	Captured Flips
AUROC	0.914	—	—
AUPRC	0.642	—	—
Sensitivity	—	10%	87.4%
Sensitivity	—	5%	71.2%
Clinically Meaningful Rate	24.8%	Top 10%	Enriched review

scoring were held constant. It also tested whether response digests and canonicalization could distinguish harmless formatting variation from clinically relevant answer changes.

The fifth experiment trained and tested the discrepancy classifier. The training data consisted of audit records and disagreement labels from the first four experiments. The test split held out complete case identifiers. Performance was evaluated by area under the receiver operating characteristic curve, area under the precision-recall curve, sensitivity at fixed review budgets, and enrichment of true disagreements among flagged cases. A useful classifier should place reproducibility failures near the top of the review queue.

The evaluation metrics included strict accuracy, macro accuracy by modality, refusal rate, nonresponsive rate, average response length, macro expected calibration error, score-flip rate, answer-flip rate, earliest differing audit layer, and reproducibility score [11]. Strict accuracy counted only exact normalized answer agreement as correct. A lenient metric allowed clinically equivalent adjacent severity terms when reviewers judged the distinction nonessential. The strict metric was primary because reproducibility should be tested under stable scoring rules.

All model calls used deterministic decoding settings where available. For hosted interfaces, a seed parameter was used only if the interface exposed one. If no seed was available, the request manifest recorded that determinism could not be guaranteed. Temperature was set to zero or the closest supported value. Maximum response length was fixed at 120 tokens. The prompt asked for a direct answer and brief justification only when necessary. The same prompt content was used across models, with interface-specific formatting recorded in the request hash.

Preprocessing decisions were documented. Radiographs were converted to 8-bit or 16-bit model-compatible images through a fixed windowing rule. CT snapshots used specified window center and width when available; otherwise a default anatomical window was applied according to task category. Ultrasound frames were converted to RGB only when required by the model interface. Dermatology photographs preserved color profiles under trace-locked evaluation and used default library behavior under ordinary logging. These decisions created observable differences in the first experiment and controlled them in the second.

Scoring was performed at the normalized answer-concept level. The rule-assisted scorer mapped parsed responses to ontology concepts and compared them with references. The language-model judge saw the question, reference concept description, model response, and a constrained rubric. It returned a correctness decision and normalized answer label. When the two scorers disagreed, the primary score used the rule-assisted scorer in pinned experiments, while disagreement was logged. Human review was used to estimate which scorer was more clinically aligned in disagreement cases.

Calibration estimation used a post hoc confidence model trained from response and interface features. The same confidence model was used across repeated runs within a condition. Confidence features included response length, certainty markers, normalized answer margin from the parser, judge confidence when available, and model interface type. Since some models did not provide logits, the confidence estimator relied on observable outputs. Calibration reproducibility was therefore partly affected by output wording and parsing.

The experiments were repeated twice for the full benchmark under each regime. Hosted-interface date-shift experiments added three temporal replicates. In total, 691,200 model responses were attempted across all experiments before excluding repeated failed requests. Persistent request failures represented 0.8% of attempts and were treated as nonresponsive for operational metrics. For reproducibility comparisons requiring paired outputs, pairs with missing output in one run were classified as failure disagreements and analyzed separately.

Statistical uncertainty was estimated by bootstrapping case

Table 9 Audit Layers in the Deterministic Evaluation Chain

Layer	Recorded Object	Hash Target	Main Purpose
Content	Raw image and question	Input bytes	Stable identity
Transformation	Processed image	Output bytes	Detect preprocessing drift
Request	Serialized payload	Prompt manifest	Verify inference input
Response	Raw and canonical text	Output digest	Detect endpoint variation
Scoring	Ontology and rubric	Score fingerprint	Stable evaluation

Table 10 Reproducibility Score Across Representative Conditions

Condition	$\mathcal{R}$ Score	Stability Level	Main Weakness
Ordinary Logging	0.781	Moderate	Hidden drift
Trace Locked	0.943	High	Hosted variation
Radiography + Gamma	0.989	Very high	Minimal residual noise
Dermatology + Beta	0.884	Moderate-high	Language variability

identifiers. For aggregate metric differences, 1,000 bootstrap replicates were used. For score-flip rates, confidence intervals were computed by clustered bootstrap over cases and model interfaces. For earliest-layer proportions, multinomial bootstrap intervals were used. For classifier evaluation, confidence intervals came from five held-out case splits. Statistical significance was not the main emphasis; effect sizes and reproducibility diagnostics were more informative.

Manual review of disagreements was conducted after automatic analysis. Reviewers inspected 1,200 disagreement cases selected from score flips, answer flips without score flips, calibration-bin boundary changes, and parser uncertainty extremes. They assigned a source label when possible: clinically meaningful model-output change, harmless wording variation, preprocessing-caused visual difference, parser error, scorer drift, or ungradable. These labels were compared with audit-derived earliest-layer labels. The comparison assessed whether audit records pointed reviewers toward the correct source.

5. Results

Under ordinary logging, cross-runtime variation was substantial. Across model interfaces and modalities, strict accuracy differed by a mean of 3.8 percentage points between nominally identical runs in different environments. The largest observed difference was 6.4 percentage points for ultrasound frames evaluated through one hosted interface. Macro expected calibration error differed by a mean of 0.117. Score-flip rate across runtime pairs was 8.9%, and normalized answer-flip rate was 11.6%. Aggregate metrics therefore understated the amount of instance-level disagreement because some flips canceled each other out.

Trace-locked evaluation reduced variation sharply. Mean strict-accuracy disagreement across runtimes fell from 3.8 to 0.6 percentage points. Macro expected calibration error disagreement fell from 0.117 to 0.021. Score-flip rate fell from 8.9% to 2.1%, and answer-flip rate fell from 11.6% to 3.4%. The remaining disagreement came mostly from hosted-interface output variation and borderline scoring cases. Local open-model runs under trace locking were almost fully repeatable, with score-flip rate below 0.4%. Hosted interfaces remained less reproducible because exact backend state could not always be verified.

The largest trace-locking gains came from image preprocessing control. Under ordinary logging, 6.7% of cases had processed-image byte differences across runtimes despite using nominally the same preprocessing script. The rate was 11.2% for dermatology photographs, 8.9% for CT snapshots, 5.4% for ultrasound frames, and 3.1% for radiographs. Dermatology differences were often caused by color profile handling. CT differences came from windowing defaults and bit-depth conversion. Ultrasound differences came from grayscale-to-RGB conversion and resizing interpolation. Under trace-locked preprocessing, processed-image byte differences were eliminated except for 0.2% of cases with corrupted metadata fallback.

Prompt and request differences were less frequent but still important. Ordinary logging produced request-hash differences in 4.3% of case-model pairs across environments. Most were caused by message-role ordering, trailing newline differences, or interface-specific serialization of image attachments. These differences produced score flips in 17.5% of affected pairs. Under deterministic request serialization, request differences fell below 0.1%. The result suggests that prompt identity should be verified at the serialized request level, not only by saving a template file.

Scoring variation alone caused meaningful metric shifts [12]. When the same stored outputs were rescored under unpinned scorer updates, strict accuracy changed by 1.9 percentage points for the rule-assisted scorer and 2.7 percentage points for the language-model judge. The judge update changed refusal interpretation and partially correct localization grading. The matcher update changed synonym mappings for severity terms and dermatologic descriptors. Pinned scoring reduced rescoring disagree-

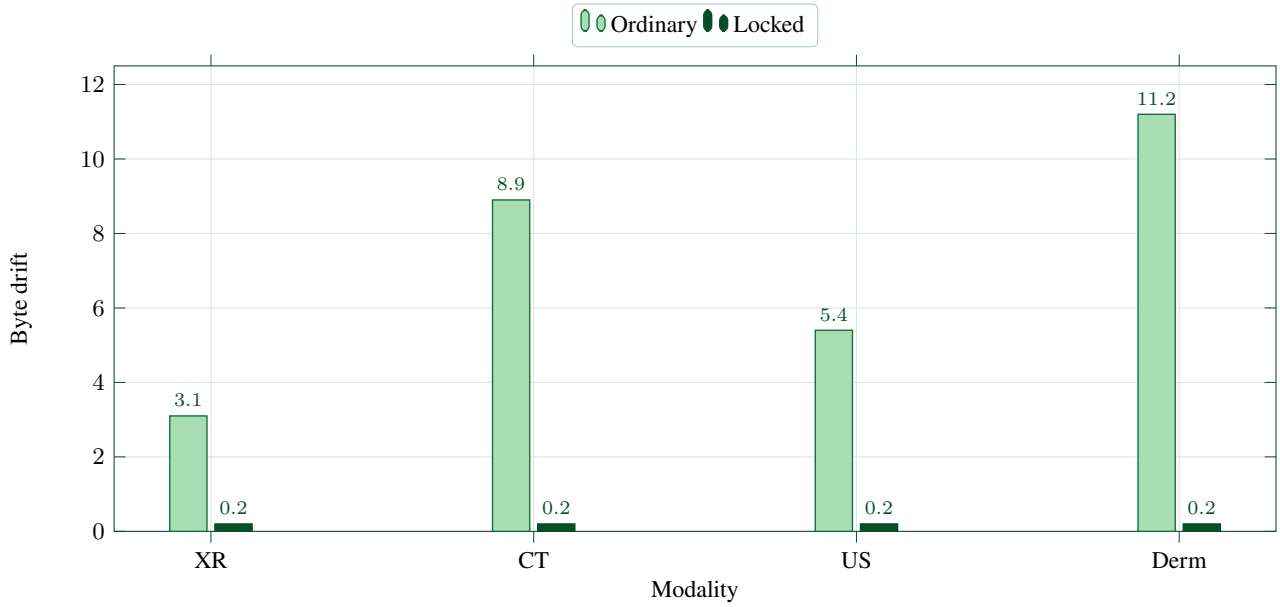

Figure 9 Processed-image byte differences were largest in dermatology and CT under ordinary logging, while trace-locked preprocessing reduced residual byte drift to the metadata fallback level.

ment to 0.3 percentage points. This experiment demonstrates that benchmark results can change even when model outputs remain fixed.

The language-model judge was more sensitive to version drift than the rule-assisted scorer. Under pinned conditions, the judge agreed with itself across repeats in 98.6% of cases. Under unpinned conditions, agreement fell to 94.1%. Most disagreements involved partially correct answers, uncertain language, and broad anatomical location. In manual review, the updated judge was clinically preferable in some cases and worse in others. The point is not that one version was wrong. The point is that unrecorded scorer changes can alter conclusions about model performance.

Hosted model interfaces showed residual variability even under trace locking. Interface Alpha had score-flip rate of 1.7% across three dates. Interface Beta had 3.9%. Local Interface Gamma had 0.3%, and local Interface Delta had 0.5%. For hosted interfaces, raw-response digest differences were common, but many were harmless formatting changes. Canonical-response digest differences were less common and more predictive of score flips. Interface Beta sometimes changed wording from absent to not clearly seen or from mild to small, which affected strict scoring in some cases. Without response digests, such changes would be difficult to distinguish from preprocessing or scoring differences.

Modality influenced reproducibility. Dermatology had the highest ordinary-logging score-flip rate at 11.8%, followed by ultrasound at 10.4%, CT at 9.1%, and radiography at 5.7%. Trace locking reduced these to 2.9%, 2.4%, 2.2%, and 1.2%, respectively. Dermatology remained highest because hosted models varied in descriptive language for color and border patterns. Radiography was most stable because grayscale preprocessing and answer ontology were simpler. CT improved strongly after windowing control but retained some scorer disagreement in small hemorrhage and mass-effect questions.

The earliest-layer analysis identified the first source of differences under ordinary logging. Among score-flip cases, 38.6% had transformation differences as the earliest detected layer, 12.4% had request differences, 24.9% had output differences despite matching transformation and request, 9.7% had parsing differences, and 14.4% had scoring differences. Under trace locking, transformation and request differences nearly disappeared. Remaining flips were 61.2% output-level, 15.8% parsing-level, and 23.0% scoring-level. This shift is expected: once controllable preprocessing and request differences are removed, irreducible model-interface and scorer issues become more visible.

Manual review confirmed the usefulness of earliest-layer labels. In 81.3% of reviewed disagreement cases, the audit-derived earliest differing layer matched the reviewer-assigned source category. Transformation-layer labels were most accurate at 91.6%. Request-layer labels matched 87.9%. Output-layer labels matched 78.4%, mainly because some output differences were clinically harmless. Parsing-layer labels matched 73.2%. Scoring-layer labels matched 75.6%. The lower parsing and scoring agreement reflects the difficulty of separating parser boundary errors from scorer interpretation errors.

The discrepancy classifier performed well. On held-out case identifiers, it achieved area under the receiver operating characteristic curve of 0.914 and area under the precision-recall curve of 0.642 for predicting score flips. Since score flips were relatively uncommon under trace locking, precision-recall performance was the more informative metric. At a review budget of 10% of case-model pairs, the classifier captured 87.4% of true score flips. At a 5% review budget, it captured 71.2%. The strongest predictors were parser uncertainty, canonical-response digest mismatch, processed-image hash mismatch, scorer margin, and hosted-interface type.

The classifier also identified clinically meaningful disagree-

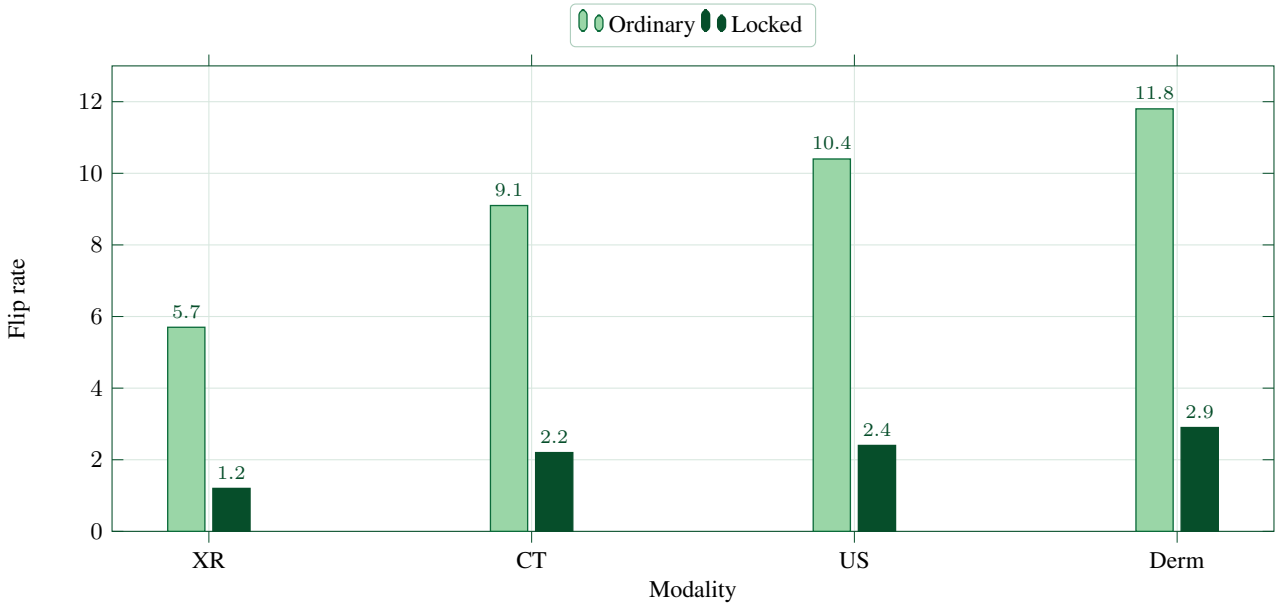

Figure 10 Score-flip rates declined for every modality after trace locking, with dermatology remaining the least stable because hosted outputs varied in descriptive language for visual attributes.

ments better than random review. Among the top 10% highest-risk cases, 24.8% contained clinically meaningful disagreement according to manual-review labels. In a random 10% sample, the rate was 6.9%. Clinically meaningful disagreements included changed laterality, changed presence or absence, changed device-position adequacy, and changed severity category. Harmless disagreements included punctuation changes, restated uncertainty with same answer concept, and equivalent synonyms. This enrichment suggests that audit-guided review can reduce manual workload.

Response length was a reproducibility signal. Cases with high response-length variability across repeats were more likely to experience score flips. The mean length difference among stable cases was 5.2 tokens, while among score-flip cases it was 21.7 tokens. Hosted interfaces produced the largest length variability. However, length variability alone was not sufficient. Some models produced different long explanations while preserving the same answer. Parser uncertainty and canonical answer mismatch were more specific predictors.

Calibration metrics were more fragile than accuracy. Under ordinary logging, calibration-bin membership changed in 14.3% of case-model pairs across runtimes. Under trace locking, this fell to 4.1%. A small number of score flips near high-confidence bins caused visible changes in expected calibration error. The worst case was a hosted interface on dermatology cases, where unpinned scoring increased expected calibration error by 0.162 despite changing strict accuracy by only 2.1 percentage points [13]. This finding shows that calibration reports require especially careful scorer and confidence-estimator versioning.

The reproducibility score $\mathcal{R}$ increased from 0.781 under ordinary logging to 0.943 under trace locking. Local interfaces achieved 0.982. Hosted interfaces achieved 0.916. The score was lowest for dermatology with hosted Interface Beta at 0.884. The score was highest for radiography with local Interface Gamma at 0.989. Although $\mathcal{R}$ is not proposed as a universal metric, it summarized the pattern: reproducibility improved when content, transformation, request, response, parser, and scorer objects were traceable.

A clinically relevant example involved CT windowing. In one runtime, a head CT snapshot was converted with a soft-tissue window; in another, a brain window was applied. The model output changed from no obvious acute hemorrhage to small hyperdense focus suspicious for hemorrhage. The reference answer was absent, so the score flipped. Under ordinary logging, both runs appeared to use the same case and script. The trace audit revealed different processed-image hashes and window parameters. Manual review found that the soft-tissue version reduced contrast in the relevant area. This was not a model reproducibility issue; it was a preprocessing identity issue.

Another example involved scoring drift. A radiograph response said the enteric tube projects below the diaphragm, likely within the stomach. The older scorer mapped this to appropriate position, while the updated scorer marked it partially correct because the tip was not explicitly named. The model output was identical. The score changed because the scorer's interpretation of sufficiency changed. The audit record identified equal response digest but differing scorer hash. This prevented the change from being misattributed to the model.

A third example involved hosted endpoint variation. The same dermatology image and request produced irregular border and color variation in one run, and mild color variation without clear border irregularity in another. Preprocessing and request hashes matched. The raw and canonical response digests differed. The score changed from correct to partially correct. No provider-exposed model version changed. The audit could not reveal the hidden backend cause, but it localized the difference to inference. This is the type of residual uncertainty that remains with opaque hosted systems.

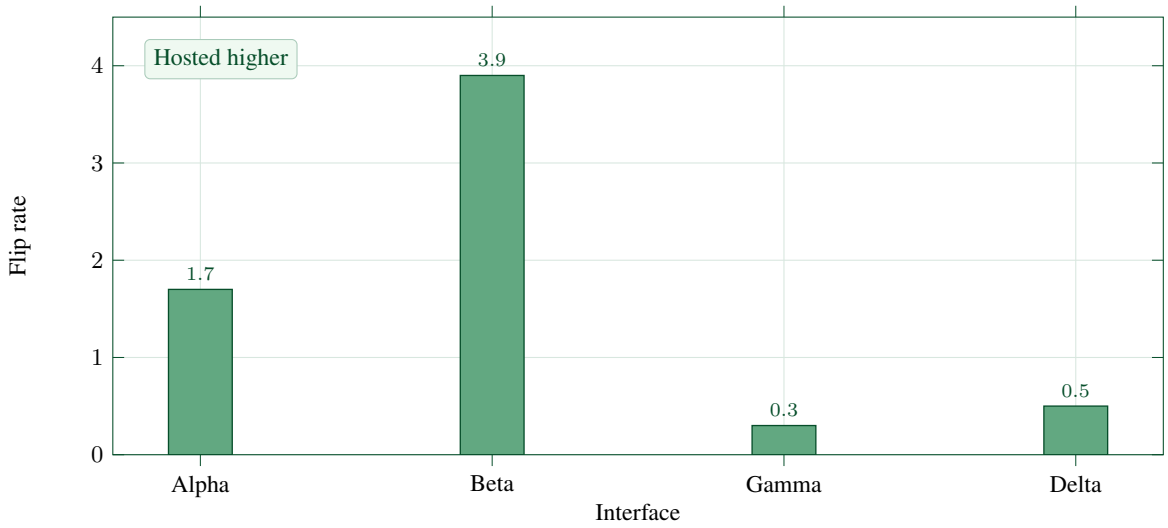

Figure 11 Residual score flips under trace locking were higher for hosted interfaces than for local interfaces, reflecting endpoint variation that remained after preprocessing, request serialization, and scoring were controlled.

The lenient correctness metric showed lower reproducibility error than strict scoring but did not eliminate disagreements. Under ordinary logging, lenient accuracy differed by 2.4 percentage points across runtimes, compared with 3.8 for strict accuracy. Under trace locking, lenient disagreement fell to 0.4 percentage points. Strict metrics are more sensitive to small wording and ontology differences, but lenient metrics can hide clinically relevant changes. The paper therefore reports both, with strict reproducibility as the main test.

6. Discussion

The results show that medical vision-language model evaluation can vary for reasons that are not visible in ordinary experiment summaries. The same dataset name, model label, prompt template, and scoring description do not guarantee the same evaluation. Image preprocessing, request serialization, model-interface behavior, parsing, and scoring all contributed to measured disagreement. The observed 3.8 percentage point accuracy variation under ordinary logging is large enough to alter model comparisons in many benchmark tables. Trace-locked evaluation reduced that variation, suggesting that much of it was preventable.

The most important practical result is that preprocessing identity must be verified at the output-byte level. It is not enough to say that images were resized to a certain resolution or converted to RGB. Different libraries can implement interpolation differently. Color profiles can be handled or ignored. CT window defaults can shift. High-bit-depth grayscale conversion can differ. These changes can alter model answers. Medical image benchmarks should therefore store or hash the exact model-facing image representation, not only the raw image and preprocessing script.

Prompt and request identity also need stronger treatment. Researchers often save prompt templates but not serialized request payloads. A template can be assembled differently across environments, and multimodal message formats can vary by interface. The request hash in this study caught differences that were otherwise easy to miss. For hosted systems, request identity is one of the few controllable elements. If the request differs, output differences cannot be interpreted as model variation alone.

Scoring reproducibility deserves the same attention as inference reproducibility. Automatic scoring is now common because generated answers are too numerous for full manual review. Yet scorers are themselves software systems, and language-model judges can change. In this study, rescoring fixed outputs with unpinned scorer versions changed accuracy by up to 2.7 percentage points. That is not a minor bookkeeping issue. A benchmark score is partly a property of the scorer. Published evaluations should therefore identify scorer versions, rubric hashes, ontology hashes, and judge prompts.

The residual variability of hosted interfaces is a difficult problem. Trace locking can ensure that the same input request was sent and that the same scoring procedure was used. It cannot fully identify hidden backend changes when the provider does not expose exact model snapshots. This does not mean hosted models cannot be evaluated. It means that reproducibility claims should be bounded. If exact backend identity is unavailable, the audit trail should say so. Repeated temporal evaluation may be needed to estimate endpoint stability.

Instance-level disagreement is more revealing than aggregate disagreement. Aggregate accuracy can remain similar while different cases flip. In medical evaluation, this matters because the identity of failure cases can be clinically meaningful. A model that alternates between correct and incorrect answers on device-position questions is not equivalent to one that has stable performance with the same average accuracy. The audit trail allowed the study to quantify case-level flips and identify their sources. Future benchmarks should report instance stability when repeated runs are feasible.

Calibration reproducibility was especially fragile. Calibration metrics depend on correctness labels and confidence estimates. If a scorer changes a small set of high-confidence answers, ex-

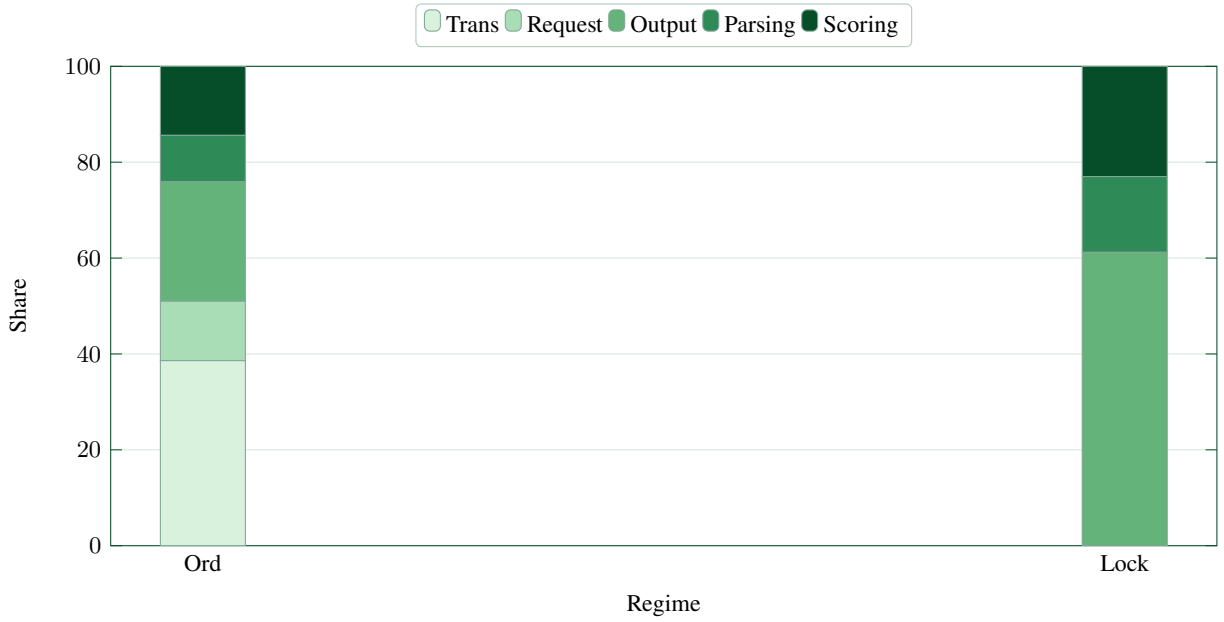

Figure 12 Earliest-layer attribution shifted after trace locking: controllable transformation and request differences nearly disappeared, leaving output, parsing, and scoring as the dominant residual sources.

pected calibration error can shift even when accuracy changes modestly. This is important because calibration is often used to justify selective review or abstention. An unstable calibration estimate can lead to unstable deployment thresholds. Calibration reports should therefore include scorer and confidence-estimator provenance, and ideally confidence-bin stability.

The discrepancy classifier suggests that audit records can support efficient review. Manual review is expensive, especially in medical imaging. By using audit metadata and output features, the classifier captured most score flips within a small review budget. This does not replace expert review. It helps choose what to review. Cases with processed-image differences, parser uncertainty, canonical-output mismatch, and low scorer margin deserve priority. A benchmark that logs these features can be audited more intelligently.

The study also clarifies the relationship between determinism and reproducibility. Deterministic decoding settings help, but they are not sufficient. A deterministic model receiving different processed images will produce legitimately different outputs. A deterministic model scored by a changing scorer will appear to change. A hosted model with deterministic parameters may still vary if the backend changes. Reproducibility requires control and identification across the full pipeline, not just temperature set to zero.

The proposed audit trail is intentionally strict [14]. Some teams may find it burdensome to hash every object and record every layer. The burden is real, but it is smaller than the cost of irreproducible medical claims. The audit schema can be automated once implemented. It does not require sharing private data publicly. Hashes and manifests can often be shared without exposing image content, although even metadata should be handled carefully. For controlled-access benchmarks, full audit records could be available to approved reviewers.

The study has implications for model leaderboards. A leaderboard that accepts only final predictions and reports a single score may not detect preprocessing or scoring drift. A stronger leaderboard would require submitted processed-input fingerprints, prompt manifests, response digests, and scorer version agreement. If the leaderboard itself performs inference, it should publish its own audit schema. If participants perform inference externally, the leaderboard should validate request and scoring provenance as much as possible.

Clinical governance also benefits from audit trails. If a hospital evaluates a model before procurement or integration, it needs to know whether the result can be repeated after software updates. If performance changes, the hospital should be able to determine whether the model changed, preprocessing changed, scoring changed, or the dataset changed. An audit trail supports this investigation. Without it, teams may spend time debating model behavior when the cause is a library update or answer ontology revision.

The findings do not imply that every evaluation must be perfectly deterministic. Some models and interfaces are inherently stochastic or opaque. The point is to record enough information to distinguish controlled variation from uncontrolled variation. If a model is nondeterministic, repeated runs should estimate variability. If an endpoint is opaque, the limitation should be reported. If a scorer is updated, old and new scores should be linked rather than silently replaced. Reproducibility is not the absence of all variation; it is the ability to identify and quantify variation.

7. Implementation Guidance

The first recommendation is to assign content-derived identifiers to every evaluation instance. These identifiers should depend on image bytes, canonical question text, reference answer concept,

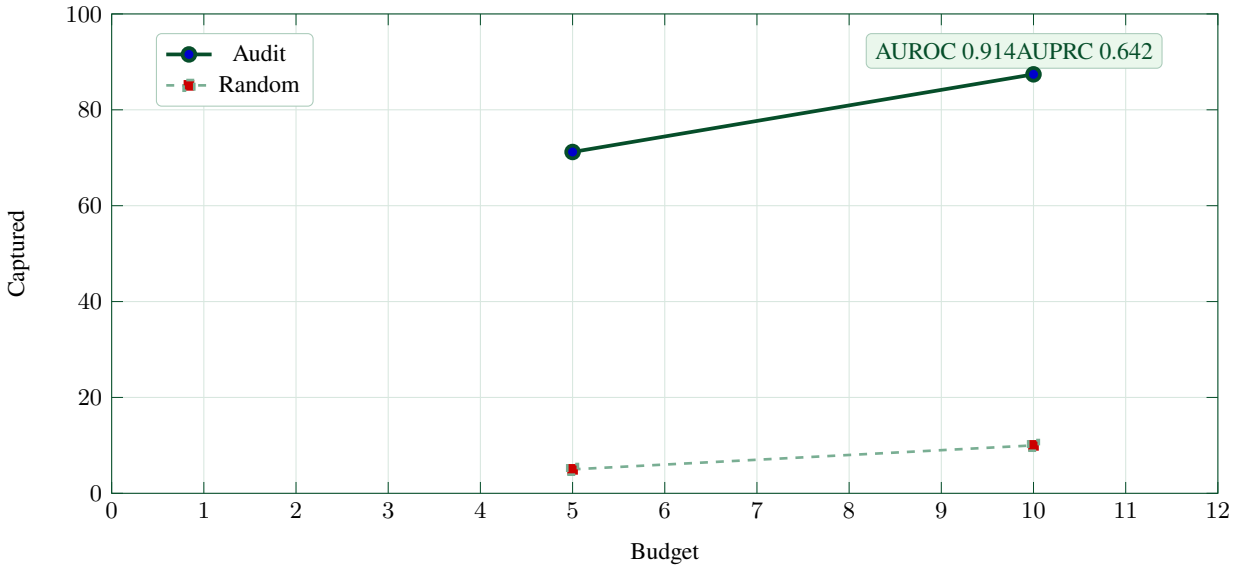

Figure 13 The discrepancy classifier captured 71.2% of true score flips at a 5% review budget and 87.4% at a 10% review budget, with AUROC 0.914 and AUPRC 0.642 on held-out cases.

and modality. They should not depend on file path or row number. File paths change easily. Row numbers change when datasets are sorted or filtered. Content-derived identifiers make accidental mixing easier to detect.

The second recommendation is to store or hash model-facing images. Raw clinical image files are important, but the model does not necessarily see the raw file. It sees a converted, resized, windowed, color-managed, compressed, or serialized object. That object should be fingerprinted. If storage permits, the exact model-facing representation should be archived in a controlled environment. If storage does not permit, its hash and preprocessing manifest should be recorded at minimum.

The third recommendation is to separate prompt templates from request payloads. A prompt template is not the same as a request. The request includes message roles, image attachment structure, serialization, decoding settings, and sometimes hidden defaults. Deterministic request serialization should be used when possible. The request hash should be recorded before submission. This allows evaluators to know whether two model calls were truly identical from the user's side.

The fourth recommendation is to keep raw and canonical response digests. Raw response equality detects exact repeatability. Canonical response equality detects clinically irrelevant formatting changes. Both are useful. A response that differs only by whitespace should not be treated the same as one that changes from present to absent. Canonicalization rules should be fixed and versioned rather than improvised after results are seen.

The fifth recommendation is to version the answer ontology. Many medical VLM evaluations depend on mappings from generated language to normalized answer concepts. These mappings are part of the measurement instrument. If they change, the score can change. Ontology files should be hashed, and changes should create a new evaluation version. Old scores should remain linked to the ontology that produced them.

The sixth recommendation is to pin automatic scorers. If a language-model judge is used, the judge prompt, model label, endpoint version if available, decoding settings, and output schema should be recorded. If exact judge backend identity is unavailable, repeated scoring should estimate judge variability. Rule-based scorers should record code, dictionary, and ontology hashes. Scoring should be treated as a reproducible computation, not a subjective afterthought.

The seventh recommendation is to report instance-level stability when possible. Aggregate score replication is useful but incomplete. Evaluators should report score-flip rate, answer-flip rate, refusal-flip rate, and modality-specific disagreement. If repeated inference is too expensive for the full benchmark, a stratified reproducibility subset can still reveal important instability. The subset should include difficult modalities, borderline answer categories, and cases with known preprocessing sensitivity.

The eighth recommendation is to log parser uncertainty. Many generated responses are not neatly formatted. A parser may struggle with hedged answers, multiple answer candidates, or conflicting statements. Parser uncertainty predicted reproducibility failures in this study. A benchmark should not treat parsing as invisible. Low parser margin cases should be reviewed or at least reported separately.

The ninth recommendation is to distinguish endpoint variation from evaluation variation. If preprocessing, request, parsing, and scoring are identical but a hosted model returns different outputs, the audit trail should identify this as inference-level variation. That variation may be acceptable or unacceptable depending on use. It should not be misattributed to dataset or scoring differences. Hosted model evaluations should include dates and endpoint metadata.

The tenth recommendation is to use audit-guided manual review. Human review resources should focus on high-risk discrepancies: score flips, low scorer margins, canonical response differences, parser uncertainty, and transformation mismatches.

Random review remains useful for estimating overall quality, but audit-guided review is more efficient for diagnosing reproducibility problems [15]. The discrepancy classifier in this study provides one way to prioritize cases.

The eleventh recommendation is to treat reproducibility as a reported metric. A paper or internal evaluation should not only say what accuracy was achieved. It should say how stable the result was under repeated runs or explain why repeated runs were not possible. A reproducibility score, score-flip rate, or audit-completeness measure can accompany ordinary performance metrics. This helps readers judge whether small differences between models are meaningful.

The twelfth recommendation is to preserve audit records for future model updates. Medical AI systems change. When a model is updated, evaluators need to compare new and old behavior. Without old audit records, it may be impossible to know whether a difference comes from the model, preprocessing, prompt, or scorer. Longitudinal audit trails support post-update validation and rollback decisions.

The thirteenth recommendation is to design benchmarks with reproducibility stress cases. Cases with high-bit-depth images, color profiles, CT windows, ultrasound overlays, borderline severity labels, and ambiguous generated answers reveal pipeline fragility. A benchmark composed only of easy, uniformly formatted images may produce reassuring but incomplete reproducibility estimates [16]. Stress cases should be labeled as such and included in repeated-run analysis.

The fourteenth recommendation is to avoid silent fallback behavior. If image conversion fails and a default path is used, the event should be recorded. If a model request times out and is retried, the retry should be recorded. If a scorer cannot parse an answer and uses a fallback judge, that should be recorded. Silent fallback is a major enemy of reproducibility because it changes computation without changing the apparent method.

The fifteenth recommendation is to make audit records machine-readable. A text methods section cannot support detailed difference localization by itself. Audit manifests should be structured, validated against a schema, and linked to each output. This allows automated comparison across runs. A human-readable summary is still useful, but the primary record should be computable.

8. Limitations

The study used a retrospective benchmark and controlled experimental conditions. Real-world evaluations may involve more varied infrastructure, more models, more institutions, and more informal handling of data. The results may therefore underestimate or overestimate reproducibility problems in particular settings. The main claim is not that the exact percentages will generalize. The claim is that reproducibility errors are measurable, nontrivial, and reducible with structured audit trails.

The benchmark used selected image-question tasks rather than full radiology report generation. Full reports would likely introduce additional reproducibility challenges, including longer outputs, section parsing, comparison statements, recommendation language, and more complex scoring. The audit formalism can extend to full reports, but the empirical results should not be assumed to apply directly. Report generation may have higher output variation and scoring drift.

The study included CT snapshots rather than full volumetric CT interpretation. Full-volume handling would add slice selection, series selection, three-dimensional preprocessing, and memory-management issues. These steps create additional audit layers. A future study should test volumetric reproducibility directly [17]. The present CT results capture only part of the problem [18].

Hosted model interfaces were treated as black boxes when exact backend versions were not available. This reflects practical reality, but it limits causal interpretation. When a hosted output changed despite identical request hashes, the audit trail could localize the difference to inference but could not identify the provider-side cause. Stronger reproducibility would require providers to expose immutable model snapshots or detailed version identifiers.

The scoring engines were only two examples. Other benchmarks may use different judges, human scoring workflows, structured extractors, or task-specific metrics. Scoring variation may be smaller or larger depending on the task. The broader point remains that scorers must be versioned and audited. Future work should evaluate more scoring strategies, including multi-expert human panels and structured clinical ontologies.

The discrepancy classifier was trained on the experimental conditions in this study. It may not transfer directly to new benchmarks, modalities, or model interfaces. Its performance should be viewed as evidence that audit features contain predictive signal, not as a universal classifier. A new evaluation program should train or recalibrate its own discrepancy model using local repeated-run data.

The audit trail uses cryptographic hashes, but hashes do not solve all privacy concerns. A hash of sensitive content is usually safer than the content itself, but metadata patterns and linkage can still create risk in some contexts [19]. Audit manifests should be governed carefully, especially when shared across institutions. Some fields may need redaction, access controls, or secure enclaves. The paper does not claim that audit logging is automatically privacy-preserving.

The study did not evaluate all possible sources of nondeterminism. Hardware-level numerical differences, batch ordering, GPU kernels, random seeds, and concurrency effects can matter for local models. The local interfaces in this study were relatively controlled. Larger distributed inference systems may show additional variation. The audit schema can record such factors, but the experiments did not exhaust them.

The manual review sample was limited. Although it was stratified and adjudicated, it covered only a subset of disagreements and outputs. Some automatic source labels may be wrong. More extensive review would improve estimates of clinically meaningful disagreement. However, the observed patterns were strong enough to support the main conclusions.

The reproducibility score $\mathcal{R}$ is descriptive and depends on chosen weights. It should not be treated as a universal standard. Different evaluations may care more about aggregate stability, instance stability, calibration, or audit completeness. The score

is useful mainly to show that reproducibility can be summarized. More work is needed to define community standards for acceptable reproducibility in medical VLM benchmarking.

9. Conclusion

This paper presented an empirical study of deterministic audit trails for reproducible medical vision-language model evaluation. Across a multi-modality benchmark, nominally identical evaluations produced materially different results when ordinary logging was used. Differences arose from image preprocessing, request serialization, model-interface variation, parsing uncertainty, and scorer version changes. These sources were often invisible in ordinary experiment summaries.

The proposed trace-locked protocol substantially reduced disagreement. By hashing image content, processed model-facing inputs, serialized requests, raw and canonical responses, parser records, scorer versions, and answer ontologies, the evaluation became more inspectable. Accuracy disagreement across runtimes fell from several percentage points to less than one percentage point, and calibration disagreement also decreased. Remaining variation was mainly attributable to hosted-interface behavior and borderline scoring decisions [20].

The study also showed that reproducibility should be measured at the instance level. Aggregate scores can hide case-level flips. A benchmark may appear stable while individual answers change. Score-flip rate, answer-flip rate, earliest differing audit layer, and audit completeness provide information that ordinary accuracy does not. The discrepancy classifier further showed that audit features can prioritize manual review of unstable cases.

The practical implication is that medical vision-language evaluation needs infrastructure discipline. Model-facing images should be fingerprinted. Requests should be serialized deterministically. Responses should be stored in raw and canonical form. Scorers and answer ontologies should be pinned. Hosted-interface evaluations should report temporal stability and endpoint uncertainty. These practices do not guarantee clinical validity, but they make benchmark numbers more trustworthy.

Reproducibility is not a decorative property of an evaluation. It is part of the measurement itself. When medical vision-language systems are compared for research, procurement, or governance, small score differences should not be interpreted without knowing whether the evaluation pipeline can reproduce them. Deterministic audit trails provide a practical way to make that question answerable.

Acknowledgments

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

References

[1] J. Guler, M. C. Roberts, M. E. Medina-Mora, R. Robles, O. Gureje, J. W. Keeley, C. S. Kogan, P. Sharan, B. Khoury, K. M. Pike, M. Kulygina, V. Krasnov, C. Matsumoto, D. J. Stein, Z. Min, T. Maruta, and G. M. Reed, ‘‘Global collaborative team performance for the revision of the international classification of diseases: A case study of the world health organization field studies coordination group.,’’ *International journal of clinical and health psychology : IJCHP*, vol. 18, pp. 189–200, 8 2018.

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

[3] J. Kennedy, K. M. Astroth, W. M. Woith, N. L. Novotny, and S. Jenkins, ‘‘New nurse graduates and rapidly changing clinical situations: the role of expert critical care nurse mentors.,’’ *International journal of nursing education scholarship*, vol. 18, 9 2021.

[4] M. I. Miller, L. C. Shih, and V. B. Kolachalama, ‘‘Machine learning in clinical trials: A primer with applications to neurology.,’’ *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics*, vol. 20, pp. 1066–1080, 5 2023.

[5] S. Wang, M. Lin, T. Ghosal, Y. Ding, and Y. Peng, ‘‘Knowledge graph applications in medical imaging analysis: A scoping review.,’’ *Health data science*, vol. 2022, 6 2022.

[6] K. C. Kuban, T. M. O’Shea, E. N. Allred, H. Tager-Flusberg, D. J. Goldstein, and A. Leviton, ‘‘Positive screening on the modified checklist for autism in toddlers (m-chat) in extremely low gestational age newborns.,’’ *The Journal of pediatrics*, vol. 154, pp. 535–540, 1 2009.

[7] J. E. Galvin, L.-C. Chang, P. Estes, H. M. Harris, and E. Fung, ‘‘Cognitive assessment with cognivue clarity®: Psychometric properties and normative ranges in a diverse population.,’’ *Journal of Alzheimer’s disease : JAD*, vol. 100, pp. 509–523, 7 2024.

[8] R. Zandi, J. D. Fahey, M. Drakopoulos, J. M. Bryan, S. Dong, P. J. Bryar, A. E. Bidwell, R. C. Bowen, J. A. Lavine, and R. G. Mirza, ‘‘Exploring diagnostic precision and triage proficiency: A comparative study of gpt-4 and bard in addressing common ophthalmic complaints.,’’ *Bioengineering (Basel, Switzerland)*, vol. 11, pp. 120–120, 1 2024.

[9] C. Paulson, S. C. Thomas, O. Gonzalez, S. Taylor, C. Swiston, J. S. Herrick, L. McCoy, K. Curtin, C. J. Chaya, B. C. Stagg, and B. M. Wirostko, ‘‘Exfoliation syndrome in baja verapaz guatemala: A cross-sectional study and review of the literature.,’’ *Journal of clinical medicine*, vol. 11, pp. 1795–1795, 3 2022.

[10] I. Blundell, R. Brette, T. A. Cleland, T. G. Close, D. Coca, A. P. Davison, S. Diaz-Pier, C. F. Musoles, P. Gleeson, D. F. M. Goodman, M. L. Hines, M. Hopkins, P. Kumbhar, D. Lester, B. Marin, A. Morrison, E. Müller, T. Nowotny, A. Peyser, D. Plotnikov, P. Richmond, A. Rowley, B. Rumpe, M. Stimberg, A. B. Stokes, A. Tomkins, G. Trensch, M. M. Woodman, and J. M. Eppler, ‘‘Code generation in computational neuroscience: a review of tools and techniques,’’ *Frontiers in neuroinformatics*, vol. 12, pp. 68–68, 11 2018.

[11] L. Blankemeier, J. P. Cohen, A. Kumar, D. V. Veen, S. J. S. Gardezi, M. Paschali, Z. Chen, J.-B. Delbrouck, E. Reis, C. Truyts, C. Bluethgen, M. E. K. Jensen, S. Ostmeier, M. Varma, J. M. J. Valanarasu, Z. Fang, Z. Huo, Z. Nabulsi, D. Ardila, W.-H. Weng, E. Amaro, N. Ahuja, J. Fries, N. H. Shah, A. Johnston, R. D. Boutin, A. Wentland, C. P. Langlotz, J. Hom, S. Gatidis, and A. S. Chaudhari, ''Merlin: A vision language foundation model for 3d computed tomography.,'' *Research square*, 6 2024.

[12] S. Badger, M.-C. Waugh, J. Hancock, S. Marks, and K. Oakley, ''Short term outcomes of children with abusive head trauma two years post injury: A retrospective study,'' *Journal of pediatric rehabilitation medicine*, vol. 13, pp. 241–253, 11 2020.

[13] C. Ding, Z. Guo, Z. Chen, R. J. Lee, C. Rudin, and X. Hu, ''Siamquality: a convnet-based foundation model for photoplethysmography signals.,'' *Physiological measurement*, vol. 45, pp. 85004–085004, 8 2024.

[14] A. P. DeLeon, ''A schizophrenic scholar out for a stroll: Multiplicities, becomings, conjurings,'' *Taboo: The Journal of Culture and Education*, vol. 17, pp. 4–, 5 2018.

[15] S. C. Harward and D. J. Englot, ''In reply: Seizure outcomes in occipital lobe and posterior quadrant epilepsy surgery: A systematic review and meta-analysis.,'' *Neurosurgery*, vol. 82, pp. 350–358, 3 2018.

[16] J. W. Gargano and M. J. Reeves, ''Sex differences in stroke recovery and stroke-specific quality of life results from a statewide stroke registry,'' *Stroke*, vol. 38, pp. 2541–2548, 8 2007.

[17] J. W. Tuke, ''Screening and surveillance of school aged children.,'' *BMJ (Clinical research ed.)*, vol. 300, pp. 1180–1182, 5 1990.

[18] C. Morgan, L. Fetters, L. Adde, N. Badawi, A. Bancale, R. N. Boyd, O. Chorna, G. Cioni, D. L. Damiano, J. Darrah, L. S. de Vries, S. C. Dusing, C. Einspieler, A.-C. Eliasson, D. M. Ferriero, D. Fehlings, H. Forssberg, A. M. Gordon, S. Greaves, A. Guzzetta, M. Hadders-Algra, R. T. Harbourne, P. Karlsson, L. Krumlinde-Sundholm, B. Latal, A. Loughran-Fowlds, C. Mak, N. L. Maitre, S. McIntyre, C. Mei, A. T. Morgan, A. Kakooza-Mwesige, D. M. Romeo, K. Sanchez, A. J. Spittle, R. B. Shepherd, M. Thornton, J. Valentine, R. Ward, K. Whittingham, A. Zamany, and I. Novak, ''Early intervention for children aged 0 to 2 years with or at high risk of cerebral palsy: International clinical practice guideline based on systematic reviews.,'' *JAMA pediatrics*, vol. 175, pp. 846–858, 8 2021.

[19] R. Hess, A. K. Santucci, K. M. McTigue, G. S. Fischer, and W. N. Kapoor, ''Patient difficulty using tablet computers to screen in primary care.,'' *Journal of general internal medicine*, vol. 23, pp. 476–480, 3 2008.

[20] Z. Su, W. Chen, S. Annem, U. Sajjad, M. Rezapour, W. L. Frankel, M. N. Gurcan, and M. K. K. Niazi, ''Adapting sam to histopathology images for tumor bud segmentation in colorectal cancer.,'' *Proceedings of SPIE–the International Society for Optical Engineering*, vol. 12933, pp. 11–11, 4 2024.