

Authority-Preserving Evaluation of Medical Vision-Language Assistants

Flint Xiaofeng Fan^{1,2}, Cheston Tan², Yew-Soon Ong², and Roger Wattenhofer¹

¹ ETH Zurich, Zurich, Switzerland

{xiafan,wattenhofer}@ethz.ch

² Agency for Science, Technology and Research (A*STAR), Singapore, Singapore

{fanx,cheston-tan,ong_yew_soon}@a-star.edu.sg

Abstract. Medical vision-language models can propose how urgently a skin lesion should be reviewed, but the local service retains authority to accept or replace that proposal under referral policy, capacity, and locally held patient context. Proposal quality and selected-action quality are therefore distinct evaluation targets, and benchmark evidence transfers between them only when local review preserves the expected action score. We introduce AuthEval, a logging and evaluation framework that records both actions, scores the selected action under declared local criteria, and, where feasible, scores the declined proposal under the same rule. It reports the resulting authority gap only when the record supports it. Because the gap is the product of the proposal-change rate and the mean score change on changed cases, that rate alone determines neither its magnitude nor its sign. On ISIC 2019, with MedGemma and simulated local review, two constraint regimes with similar change rates produced an optimistic image-equal gap under capacity (+0.744 simulator units) but no detectable gap under safety. The declared evaluation unit also mattered: under mixed constraints the gap reversed from +0.374 to −0.206 when weighting shifted from image to lesion-aware cluster. AuthEval thus clarifies whether a study’s records support claims about the model, the workflow, or both.

Keywords: Medical vision-language models · Evaluation construct validity · Clinical decision support · Human–AI workflows

1 Introduction

A lesion image enters a teledermatology pathway, and a medical vision-language model (VLM) proposes how urgently it should be reviewed. The local service, however, retains authority over the routing decision: referral policy, available capacity, follow-up safeguards, and clinician review may lead it to accept or replace the proposal [13, 9]. The action retained by the workflow is therefore the one the service selected, which need not be the one the model proposed.

Medical VLM benchmarks evaluate model outputs against references [16, 8], addressing proposal quality. Selected-action quality becomes necessary when

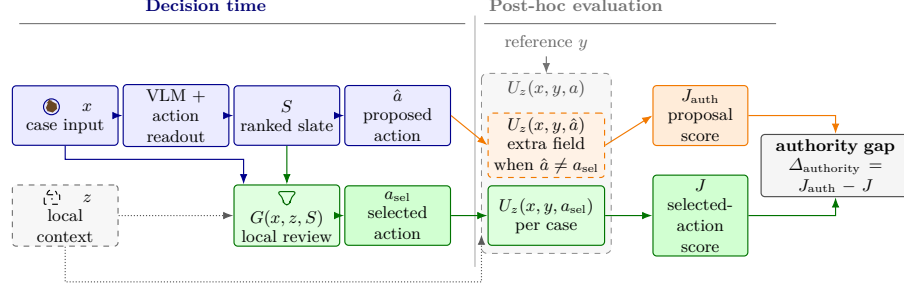


Fig. 1. Why proposal and selected-action scores can differ, organized by *when* information is available. At decision time the proposer sees only x , while the label-blind gate G also sees local context z and slate S , but never the reference y . Post-hoc scoring applies one locally parameterized rule to both actions; the declined proposal’s score is the one extra field needed when review changes the action.

the claim concerns the combined workflow, because the workflow’s routing decision is the action retained after local review, not a proposal the service rejects. Moving from proposal quality to selected-action quality changes the evaluation construct [15]. Local review may preserve the proposal, replace a weak proposal with a better-scoring action, or replace a strong proposal because local constraints make it inadmissible, so the two targets can diverge in either direction.

A role-separated log with both actions and the selected-action score reveals how often local review changes the proposal and how well the selected actions score. It is still insufficient to measure the authority gap: on changed cases, the score that the declined proposal would have received under the same local criteria remains unknown. Because this score may depend on locally held context, it must be computed where that context is available. Figure 1 summarizes the two evaluation paths and this boundary.

This study asks which recorded fields suffice to recover the authority gap, whether the frequency of local change determines it, and how it varies with the review process, the proposal mechanism, and the declared evaluation unit. We answer these with AuthEval, a logging protocol and evaluator that records the proposed and selected actions, scores both under the same local criteria when feasible, and otherwise marks the gap as not recoverable rather than treating proposal quality as workflow quality. Factorizing the gap into the proposal-change rate and the mean score change on changed cases shows why change frequency alone cannot recover it, and whether proposal scoring is optimistic or pessimistic relative to selection. In a controlled ISIC 2019 study the gap varies across all three dimensions and reverses sign between image- and cluster-equal weighting.

2 Related Work

Medical VLM benchmarks evaluate model outputs across a range of tasks [16, 8], but a different question arises when a study seeks evidence about the combined

workflow, in which a local service may accept or replace the model’s proposal. AuthEval complements proposal evaluation by asking whether proposal-level evidence remains representative after local selection.

Clinical alert studies distinguish appropriateness from provider response [10]. Human–AI research separates advice, privately held information, and final decisions [1] and distinguishes reliance frequency from its payoff [4], while learning-to-defer methods optimize which actor decides [12]. AuthEval addresses the complementary measurement problem of determining which recorded fields suffice to recover the signed proposal-selection score difference under the same local criteria.

3 Authority-Preserving Evaluation (AuthEval)

We call an evaluation *authority-preserving* when its record keeps the model’s proposal and the accountable local selection as distinct objects scored under the same declared rule. The term describes the evaluation record, not the clinic: AuthEval neither allocates clinical authority nor certifies that a local selection was appropriate.

A case provides proposer-visible information x and locally held clinical or operational context z , together with action set $\mathcal{A}$. A medical VLM and declared action readout form a policy π that produces a ranked list of options, or slate, $S_\pi(x)$. Its first option is the proposal $\hat{a}$. A label-blind local review rule, gate G , selects $a_{\text{sel}} = G(x, z, S_\pi(x))$ and may accept the proposal or substitute another listed action. Post-hoc evaluation additionally uses a fixed reference state y , a diagnosis-derived risk target used only for scoring. The gate does not observe y , so its selection cannot use the evaluation target.

Let $U_z(x, y, a)$ be a declared post-hoc action score, with larger values preferred. It scores either action using the same local criteria and fixed reference y , and all expectations use the declared weighting rule. Under this rule, $J(\pi) = \mathbb{E}[U_z(x, y, a_{\text{sel}})]$ is the selected-action score, while $J_{\text{auth}}(\pi) = \mathbb{E}[U_z(x, y, \hat{a})]$ is the local-rule proposal score. On changed cases, computing J_{auth} requires the score that the declined proposal would have received under those same criteria. This declined-proposal score is not an estimate of the patient outcome that the proposal would have caused. Their authority gap, $\Delta_{\text{authority}} = J_{\text{auth}} - J$, isolates the score change introduced by local selection. Positive values mean that proposal scoring is optimistic relative to selection, and negative values mean that it is pessimistic.

Proposition 1 (Authority-gap factorization). *Let $I = \mathbf{1}[\hat{a} \neq a_{\text{sel}}]$, $p = \Pr(I = 1)$, and, for $p > 0$, $\delta = \mathbb{E}[U_z(x, y, \hat{a}) - U_z(x, y, a_{\text{sel}}) \mid I = 1]$. Under any fixed sampling and weighting rule, $\Delta_{\text{authority}} = p\delta$. At $p = 0$, $\Delta_{\text{authority}} = 0$ and δ is undefined. Consequently, for $p > 0$, the proposal-change rate does not determine $\Delta_{\text{authority}}$ without an assumption fixing δ , and two local review processes can share p while their authority gaps differ in magnitude or sign.*

Cases with $I = 0$ contribute zero, and conditioning the remainder on $I = 1$ gives the factorization: p is the proposal-change rate, the proportion of evaluation units whose routing changes under the declared weighting rule, and δ is the average signed score difference on those changed units under the same local criteria. Because p leaves δ free, equal proposal-change rates can produce authority gaps with different magnitudes or directions. If review changes 30 of 100 proposals, then $p = 0.30$, and a mean advantage of two units for the declined proposals gives $\Delta_{\text{authority}} = +0.60$, while the same 30 changes improving the selected action by two units instead give $\Delta_{\text{authority}} = -0.60$. One change rate thus supports opposite conclusions.

Record recoverability. We call a quantity *directly recoverable* when the recorded fields suffice to compute it without additional assumptions. A record containing both actions and the selected-action score does not by itself recover $\Delta_{\text{authority}}$ when proposals change, because U_z depends on locally held z : when z remains local, the evaluation record must additionally include the declined proposal’s locally computed score under the same criteria (Section 6).

4 Experimental Setup

We instantiate AuthEval in a controlled study: proposals come from frozen models scored on real dermoscopy images, while local review and context are simulated under a study-defined rule. We use the ISIC 2019 dermoscopy training set, comprising Memorial Sloan Kettering (MSK), HAM10000, and BCN20000 sources [6, 2, 18, 5]. Diagnoses map to four study-defined risk levels, very low to high, and four ordinal routing actions span observation to urgent specialist assessment (Supplementary Material gives the mapping). Lesion-aware splitting with image fallback produces 17,629/3,735/3,967 train/validation/test images, and known-lesion or exact-pHash groups do not cross splits (a broader, unenforced pHash diagnostic at Hamming threshold 4 finds 122 cross-split near-duplicate components, detailed in the Supplementary Material). Code and reproduction instructions are available at <https://github.com/flint-xf-fan/AuthEval>.

Frozen MedGemma 1.5 4B assigns each risk continuation a mean token log-likelihood from the image and proposer-visible age, sex, and anatomic site. A validation-trained balanced multinomial mapping, the action readout, ranks the four actions from these likelihoods [17]. The readout was selected and locked before test access; its test balanced accuracy is 0.420. Earlier ISIC releases entered MedGemma pretraining [17], so “held out” applies to readout fitting, not foundation-model pretraining.

Services may cap escalation (capacity), impose a safety floor (safety), or apply both (mixed). Private factors can additionally remove or replace an admissible action and rescale the scoring weights below. A class-balanced ridge metadata proxy, fitted on training labels, supplies decision-time risk context from age, sex, and anatomic site. Four frozen synthetic profiles set per-profile capacity and

safety bounds, and the gate selects the first admissible action from the ranked slate, never observing the test-case diagnosis, ideal action, or utility score. The Supplementary Material gives the full profile assignment, private-factor mechanics, selection rule, and prompt.

With $\ell(a), \ell(y) \in \{0, 1, 2, 3\}$ the action and reference levels, the declared score is $U_z(x, y, a) = -w_u(z) s_y [\ell(y) - \ell(a)]_+ - w_o(z) [\ell(a) - \ell(y)]_+$, where $[t]_+ = \max(t, 0)$ and s_y is the component indexed by $\ell(y)$ in $(s_0, s_1, s_2, s_3) = (0.5, 1, 2, 4)$. The under- and over-escalation weights w_u, w_o have base values set per profile and are rescaled by the private factors above. Zero is best, and a one-level over-escalation costs 0.71–3.45 units versus 4.6–9.0 for under-escalation at high risk, so upward and downward substitutions are non-equivalent under the declared static score. The Supplementary Material lists the per-profile weight pairs and effective ranges.

Image-equal estimates weight images, and cluster-equal estimates weight 2,090 known-lesion/exact-pHash clusters. Confidence intervals (CIs) use 5,000 whole-cluster bootstrap replicates and are conditional on the fitted readouts and frozen proxy, gate, utility, and simulated context. For the mechanism comparison, cases, context, gate, scoring rule, and metadata proxy remain fixed while the proposal mechanism varies: MedGemma is compared with a LLaVA-Med option-likelihood control whose expected ordinal score is discretized by validation-fitted thresholds (balanced accuracy 0.249) [7, 11], and a validation-thresholded operational proposal derived from a supervised linear readout over LLaVA-Med’s frozen CLIP vision tower (0.546 balanced accuracy) [14].

5 Results

Proposal and selected-action scores diverge under local review. Under the mixed simulated local review process, the selected action differs from the MedGemma-derived proposal on 47.0% of test images. The proposal score $J_{\text{auth}} = -2.94$ and selected-action score $J = -3.31$ give an authority gap $\Delta_{\text{authority}} = +0.374$ simulator units (95% CI $[+0.091, +0.663]$): under this declared score, proposal-only scoring is optimistic by that amount. The proposal-change rate $p = 0.470$ and mean score change $\delta = 0.797$ give $p\delta = 0.374$ after rounding, consistent with Proposition 1. Scoring the declined proposal makes this comparison possible; the change rate remains simulator-specific. Table 1 summarizes this result and three checks.

Change frequency does not determine the authority gap. Capacity and safety yield proposal-change rates of 0.341 and 0.281, but mean score changes of 2.18 and -0.18 on changed cases. The corresponding image-equal authority gaps are $+0.744$ ($[+0.491, +1.010]$) and -0.051 ($[-0.262, +0.158]$); capacity substitutions are 97.6% downward, whereas the safety arm is near zero. A rate-only adjustment with one shared conditional value would give both gaps the same sign and scale them only by p . Instead, they differ by 0.795 units with opposite point-estimate signs, though only the capacity interval excludes zero. Change frequency therefore determines neither size nor direction.

Table 1. Controlled authority gaps on the ISIC test split ($n = 3,967$ images; 2,090 clusters; simulator units); positive values denote optimistic proposal scoring. The final row is the paired probe-minus-LLaVA-Med gap difference under fixed capacity. p_{img} is the image-level proposal-change rate and δ_{img} the mean score change on changed images. Weighting and bootstrap CIs (brackets) are defined in Section 4.

Quantity	Setting	p_{img}	δ_{img}	Image estimate	Cluster estimate
$\Delta_{\text{authority}}$	MedGemma / mixed	0.470	0.80	+0.374 [+0.091, +0.663]	−0.206 [−0.373, −0.037]
$\Delta_{\text{authority}}$	MedGemma / capacity	0.341	2.18	+0.744 [+0.491, +1.010]	+0.056 [−0.107, +0.218]
$\Delta_{\text{authority}}$	MedGemma / safety	0.281	−0.18	−0.051 [−0.262, +0.158]	−0.224 [−0.358, −0.093]
Paired gap	Fixed capacity	−	−	+0.810 [+0.629, +0.999]	+0.433 [+0.345, +0.524]

The gap varies across proposal mechanisms. Under a fixed capacity gate, the image-equal authority gap is +0.014 for a LLaVA-Med option-likelihood head, +0.744 for MedGemma, and +0.824 for a linear probe over LLaVA-Med’s frozen CLIP vision tower. A paired cluster bootstrap places probe minus LLaVA-Med at +0.810 (95% CI [+0.629, +0.999]), and at +0.433 (95% CI [+0.345, +0.524]) under cluster-equal weighting. Because the mechanisms differ in architecture, readout, input signals, and slate construction, this analysis supports neither component attribution nor a model-quality comparison; balanced accuracy does not explain their gap differences. A capacity ceiling binds mechanisms in proportion to how often they escalate through predominantly downward substitutions. Thus, a gap for one complete proposal mechanism need not hold for another under the same review process.

What counts as one case changes the estimate. Giving each image equal weight produces the primary MedGemma authority gap +0.374, while giving each lesion-aware cluster—images sharing a known lesion ID or exact perceptual-hash match, with image fallback—equal weight produces −0.206 ([−0.373, −0.037]). The two units differ because lesions are photographed unequally: a high-risk cluster holds 2.98 images on average against 1.54 for a very-low-risk cluster. Image weighting therefore emphasizes repeated high-risk lesions, precisely where a capacity ceiling’s downward substitution is most expensive under the asymmetric score. Cluster weighting instead gives each lesion one vote.

Capacity changes from +0.744 to +0.056 ([−0.107, +0.218]) and safety from −0.051 to −0.224 ([−0.358, −0.093]): under cluster weighting, the capacity estimate becomes indistinguishable from zero while the safety estimate excludes zero. Together, these checks show that proposal accuracy and change frequency cannot substitute for direct measurement of the authority gap, and that the proposal mechanism and evaluation unit must be declared: AuthEval enables that measurement from a full counterfactual record.

6 Operationalizing AuthEval and Limitations

Measuring the authority gap in a service is a recording problem before it is a statistical one. A proposal-only log supports evaluation of the model’s recommendation, but not of the selected workflow. Adding the action the service selected, together with its score under the local rule, shows how often review changes the proposal and how well the selected actions score, but not whether those changes improve the declared score. Answering that requires one further field, one not ordinarily produced during operation: the score the declined proposal would have received under the same local rule. Because that rule reads locally held context, the score has to be computed at the site, either from a computable policy or through retrospective adjudication that does not alter the original selection. Local computation keeps the underlying context at the site, reducing unnecessary exposure in the evaluation record [3]. Where the score cannot be obtained, AuthEval reports the gap as not recoverable rather than letting proposal quality stand in for workflow quality. The Supplementary Material states the record levels and the reporting rules exactly.

Our results come from one imaging task in which the review process, the private context, the workflow profiles, and the scoring weights are all study-defined, so they establish that an authority gap can arise and change sign rather than estimating its size in clinical practice. The intervals are conditional on those authored choices, and the three proposal mechanisms differ along multiple dimensions. The Supplementary Material details these boundaries and the evidence needed to extend the claims beyond them.

7 Conclusion

Medical VLM benchmarks measure the quality of a model’s proposed action, whereas clinical workflows also require a service to select an action under local capacity, policy, and patient constraints. Treating these as the same evaluation object can make workflow performance appear better or worse for reasons introduced after the model’s proposal. We introduced AuthEval to keep these roles separate: it records the proposed and selected actions, evaluates both under the same local criteria, and expresses their difference as the authority gap. That gap is the product of the change rate and the average shift when it happens, so frequency alone cannot recover it. In our controlled ISIC study, its magnitude and direction varied across simulated review processes, proposal mechanisms, and evaluation units, and, under mixed constraints, it reversed sign when weighting shifted from image to lesion-aware cluster. AuthEval therefore lets studies making workflow-level claims test whether proposal scores remain representative after local review, measuring the resulting gap when local rescoring is available and stating clearly which workflow claims the available record can support.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Bordt, S., von Luxburg, U.: A bandit model for human-machine decision making with private information and opacity. In: *Proceedings of AISTATS* (2022)
2. Codella, N.C.F., Gutman, D., Celebi, M.E., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). *arXiv:1710.05006* (2017)
3. Fan, F.X., Tan, C., Wattenhofer, R., Ong, Y.S.: Position paper: Rethinking privacy in RL for sequential decision-making in the age of LLMs. In: *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE (2025)
4. Guo, Z., Wu, Y., Hartline, J., Hullman, J.: A decision theoretic framework for measuring AI reliance. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (2024)
5. Hernández-Pérez, C., Combalia, M., Podlipnik, S., et al.: BCN20000: Dermoscopic lesions in the wild. *Scientific Data* (2024)
6. International Skin Imaging Collaboration: ISIC challenge datasets: 2019. <https://challenge.isic-archive.com/data/> (2019)
7. Li, C., Wong, C., Zhang, S., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: *NeurIPS DB Track* (2023)
8. Liu, C., Pan, J., Shen, W., et al.: How far have medical vision-language models come? a comprehensive benchmarking study. *arXiv:2507.11200* (2025)
9. Marsden, H., Kemos, P., Venzi, M., et al.: Accuracy of an artificial intelligence as a medical device as part of a UK-based skin cancer teledermatology service. *Frontiers in Medicine* (2024)
10. McCoy, A.B., Waitman, L.R., Lewis, J.B., et al.: A framework for evaluating the appropriateness of clinical decision support alerts and responses. *Journal of the American Medical Informatics Association* (2012)
11. Microsoft: LLaVA-Med v1.5 Mistral 7B model card. Hugging Face (2024), <https://huggingface.co/microsoft/llava-med-v1.5-mistral-7b>
12. Mozannar, H., Sontag, D.: Consistent estimators for learning to defer to an expert. In: *Proceedings of ICML* (2020)
13. National Institute for Health and Care Excellence: Artificial intelligence (AI) technologies for assessing and triaging skin lesions referred to the urgent suspected skin cancer pathway: Early value assessment. *HealthTech guidance HTG746* (2025)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. PMLR (2021)
15. Raji, D., Denton, E., Bender, E.M., Hanna, A., Paullada, A.: AI and the everything in the whole wide world benchmark. In: *NeurIPS DB Track* (2021)
16. Royer, C., Menze, B., Sekuboyina, A.: MultiMedEval: A benchmark and a toolkit for evaluating medical vision-language models. In: *Proceedings of the 7th Conference on Medical Imaging with Deep Learning* (2024)
17. Sellergren, A., Gao, C., Mahvar, F., et al.: MedGemma 1.5 technical report. *arXiv:2604.05081* (2026)
18. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions. *Scientific Data* (2018)