NOT ALL GAINS ARE EQUAL: CONTROLLED ATTRIBUTION IN MEDICAL FOUNDATION MODEL DEPLOYMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Local adaptation can improve a medical foundation model's performance, but a lower post-adaptation error alone does not reveal whether the gain comes from correcting systematic output shift, using local supervision, or changing the reliability of downstream decisions. We introduce a controlled deployment attribution framework that separates calibration-recoverable improvement, the supervision-dependent value of released predictions, and the decision-level effect of predictive uncertainty. We instantiate the framework on PanEcho using 10,175 echocardiography sequences linked to 8,283 reports with seven continuous measurements. Before crediting a replacement head with the full improvement, we find that train-only affine calibration recovers 74.5% of the specified macro normalized-MAE gap between released predictions and a historical local MLP. Released predictions are most useful as an anchor under limited local supervision, while their advantage narrows as labels increase. Propagating patient-group prediction intervals through fixed measurement-derived rules changes the downstream risk–coverage trade-off: false assertions decrease from 10.52% at 100% assertion coverage to 2.47% at 49.40% coverage. In this controlled deployment setting, the results support attributing observed local gains to calibration, supervision, and decision-level effects rather than to a single adaptation gain.

1 INTRODUCTION

Released medical foundation models provide pretrained representations and predictors for institution-specific use. Their availability enables local adaptation to clinical data, with improvements commonly assessed by comparing released predictions with those of a locally adapted model. While such comparisons quantify changes in performance, a key obstacle to interpreting adaptation gains lies in our limited understanding of what produces these changes. In echocardiography, systematic output shifts, the availability of local supervision, and the interpretation of uncertain measurements can affect different aspects of deployment. Understanding these effects is therefore important for determining when and why local adaptation is useful.

The interplay between pretrained predictors and institution-specific data makes deployment gains difficult to interpret. For example, a locally trained regression head can correct a global offset or scale mismatch while also learning case-specific corrections, making it unclear whether its improvement requires learning beyond a simple output transformation. Retaining the released prediction may also help when local labels are limited, even though this prediction is derived from the same frozen representation available to the local head. Furthermore, numeric estimates may be mapped to fixed downstream rules, where measurement uncertainty can produce unsupported definite assertions. These effects concern different properties of the deployed model, yet a single regression-error comparison does not distinguish them.

Although released-to-local comparisons reveal whether performance improves, several questions remain unresolved. Notably, they do not establish how much of the observed error gap can be recovered through output calibration alone, or how the value of retaining released predictions changes with the amount of local supervision. Furthermore, improved regression accuracy does not determine when an uncertain measurement supports a definite rule category. A systematic evaluation of these questions is needed to distinguish calibration-recoverable improvement, supervision-dependent correction, and downstream assertion reliability.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

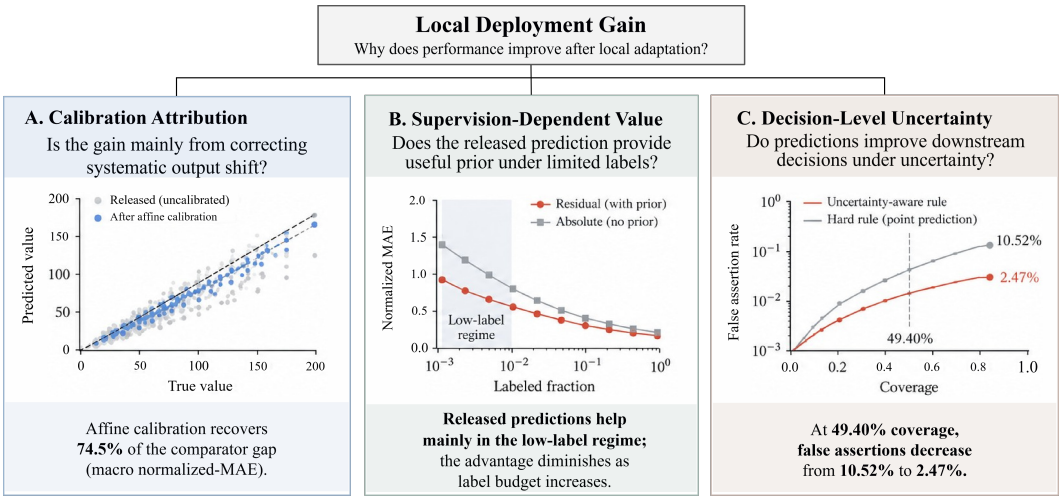


Figure 1: Three complementary axes for attributing local deployment gains.(A) Calibration attribution quantifies the recoverable released-to-local gap through train-only output calibration.(B) Supervision-dependent comparison evaluates the value of retaining released predictions across different local label budgets.(C) Decision-level uncertainty evaluates measurement-derived assertions by jointly considering false assertions and definite coverage. The three axes represent distinct and non-additive aspects of deployment gain attribution.

We introduce a controlled deployment attribution framework to investigate these questions along three dimensions (Figure 1). The first, *calibration attribution*, measures the portion of a specified released-to-local regression-error gap recovered by train-only output calibration. The second, *supervision-dependent comparison*, examines whether retaining released predictions provides additional value under different levels of local supervision by comparing absolute and released-output-anchored residual heads on matched patient subsets and label budgets. The third, *decision-level uncertainty evaluation*, investigates how measurement uncertainty affects the reliability of measurement-derived rule assertions by propagating independently calibrated intervals through prespecified rules and evaluating false assertions jointly with definite coverage. These dimensions define distinct evaluation targets rather than additive components of a single adaptation gain.

We design a series of controlled experiments using PanEcho, whose shared video representation and task-specific heads support 39 echocardiography reporting tasks (Holste et al., 2025). Our single-center cohort comprises 10,175 echocardiography sequences linked to 8,283 reports, with evaluation focused on seven strictly mapped continuous measurements. To support consistent comparisons across the three dimensions, we freeze the pretrained representation, use patient-isolated data partitions, fit output calibration on training data only, and compare regression methods on matched evaluation rows. For downstream evaluation, a separate proper-training/calibration partition provides patient-group measurement intervals for four frozen single-measurement rules. Reference categories are generated by applying the same rules to the observed numeric measurements.

Our findings are as follows.

(1) Calibration explains a substantial portion of apparent adaptation gains. Train-only affine calibration recovers 74.5% of the macro normalized-MAE gap between released predictions and the historical primary local MLP. This lightweight output transformation closes much of the specified gap without learning a new feature-based prediction head, so the full released-to-local improvement should not be attributed to local-head fitting.

(2) Released predictions provide supervision-dependent inductive bias. The benefit of released-output anchoring is most evident when local labels are limited, despite using the same frozen representation.

(3) Decision reliability requires uncertainty-aware evaluation beyond regression error. At the primary 90% interval setting, false assertion among definite outputs decreases from 10.52% at 100%

coverage to 2.47% at 49.40% definite coverage. This risk–coverage trade-off under measurement-derived rules shows that regression accuracy alone does not characterize the reliability of definite measurement-derived assertions.

Together, these findings motivate interpreting observed deployment gains through calibration, supervision, and downstream uncertainty rather than a single adaptation effect. Our conclusions concern the controlled single-center PanEcho setting, with measurement-derived rule categories rather than independently adjudicated clinical outcomes as the downstream evaluation endpoint.

2 Related Work

2.1 Evaluating Medical Foundation Model Adaptation

Deep echocardiography systems have progressed from view classification and automated measurement pipelines to video-based ventricular-function assessment (Madani et al., 2018; Zhang et al., 2018; Ouyang et al., 2020), while chamber-quantification recommendations define much of the measurement vocabulary used in clinical reporting (Lang et al., 2015). More broadly, medical foundation models aim to support reusable representations across tasks (Moor et al., 2023). PanEcho instantiates this direction for echocardiography with 39 reporting tasks, a shared image encoder, a temporal transformer, and task-specific heads over 16-frame clips (Holste et al., 2025).

Local reuse is not automatically equivalent to learning new information. Medical-imaging studies show that transfer behavior depends on the target task and fine-tuning regime (Tajbakhsh et al., 2016; Raghu et al., 2019), while cross-institution evaluation and shortcut analyses demonstrate that apparent performance can change under acquisition and site shift (Zech et al., 2018; DeGrave et al., 2021). Parameter-efficient adaptation updates restricted pretrained components (Hu et al., 2022); multiple-instance learning addresses aggregation over weakly aligned acquisitions (Ilse et al., 2018); and task weighting can balance unequal supervision (Kendall et al., 2018). In our local data, roughly 82% of reports have one matched sequence and reliable view names are unavailable, so report mean pooling and masked task losses provide simple baselines. Our work differs by holding the released encoder fixed and asking what a released-to-local performance gap actually measures, using controlled comparisons instead of treating local retraining as a single undifferentiated gain.

2.2 Calibration and Output Shift

Post-hoc calibration is well established for classifier confidence (Guo et al., 2017), and regression work extends calibration to predictive distributions and uncertainty estimates (Kuleshov et al., 2018; Song et al., 2019). These probabilistic objectives differ from calibrating a continuous point prediction, but they motivate separating correctable output mismatch from representation-dependent error. A replacement head can correct systematic offset and scale while also learning case-specific residuals, making the direct zero-shot–adapted difference ambiguous. Our work differs by quantifying the share of a fixed released-to-local gap recovered by train-only intercept and affine maps, then measuring the remaining local-head advantage on the same frozen representation and evaluation rows.

2.3 Uncertainty-Aware Decision Making

Predictive-uncertainty methods span model-based approximations, ensembles, calibration, and distribution-free constructions (Gawlikowski et al., 2023). Selective classification provides an explicit error–coverage trade-off (Geifman & El-Yaniv, 2017), while distribution-free regression and conformal prediction provide finite-sample marginal coverage under exchangeability (Lei et al., 2018; Angelopoulos & Bates, 2023). In clinical reporting, factual-completeness objectives, semantic rewards, and structured knowledge extraction address a separate generation problem (Miura et al., 2021; Delbrouck et al., 2022; Dalla Serra et al., 2022). Our work differs by treating uncertainty as a decision-level evaluation: patient-group intervals are propagated through fixed measurement rules, false assertions are reported jointly with definite coverage against a measurement-derived reference, and any optional language renderer remains downstream of deterministic rule ownership.

3 Data and Problem Formulation

Matching and split. The primary quantitative cohort contains extracted JPEG frame paths, and its local study identifiers are cohort identifiers rather than DICOM UIDs. The source contains 52,999 structured reports and exported image sequences. We retain stable sequence directories and use exact identifiers only: an eight-digit image or study identifier is matched first, followed by exact patient identifier and examination date when needed. The manifest construction records one uniformly decoded frame for availability checking; model input loading separately samples 16 decodable frames. Of 10,374 stable sequence candidates, 10,175 sequence rows match successfully (98.08% sequence-candidate match rate); they correspond to 10,039 studies and 8,283 unique reports. This rate describes sequence-candidate matching. The existing split excludes 156 matched rows without finite LVEF, the target originally used to construct the split. The remaining patient-disjoint train, validation, and test partitions contain 6,997, 1,521, and 1,501 sequences from 5,596, 1,199, and 1,199 patients and 5,720, 1,217, and 1,227 reports. Patient, study, report, and composite-sequence intersections across partitions are zero. As a sensitivity analysis, we also construct a target-independent split over all 10,175 rows and 8,110 patients without conditioning on any finite measurement. Its train, validation, and test partitions contain 7,139, 1,520, and 1,516 sequences. We extract a new isolated feature cache and retrain the strict-seven MLP once on this split as split-sensitivity evidence.

Targets. Automatic field analysis excludes identifiers, dates, categorical codes, near-constant values, and numeric metadata. The historical pipeline applies numeric coercion without a prespecified physiological-range filter; for example, the training LVEF field contains a maximum of 560. We retain the historical analysis for comparability and separately apply a frozen train-fitted robust filter, removing labels outside $[Q_1 - 3\,\mathrm{IQR}, Q_3 + 3\,\mathrm{IQR}]$ with bounds estimated on training labels only. These train-fitted statistical sensitivity bounds transfer unchanged to validation and test. Field analysis initially identifies 28 candidate fields. An exact-duplicate check finds that the field labeled pulmonary-artery peak velocity is identical to main pulmonary-artery diameter in all 10,059 jointly observed rows, including missingness. We retain the diameter and exclude the mislabeled duplicate from distinct-signal counts and future training. The corrected counts are 27 distinct candidate signals, 25 manifest-present trainable signals, 20 test-evaluable signals in the broad run, 16 recommended signals, and 15 core signals. We report these categories separately because field count, trainability, evaluability, and distinct clinical signal are different claims.

Learning problem. Let s index a sequence, $p(s)$ its patient, and $t \in \mathcal{T}$ a measurement task. A sequence comprises $F = 16$ RGB frames, $\mathbf{x}_s \in \mathbb{R}^{3\times F\times 224\times 224}$. The report supplies target y_{st} when mask $m_{st} = 1$; otherwise $m_{st} = 0$. Given patient-disjoint partitions, the sequence-level problem is to learn $\hat{y}_{st} = f_t(\mathbf{x}_s)$ while excluding missing labels from the loss. Because y_{st} originates from the examination report rather than a frame annotation, this formulation is weakly aligned: report measurements supervise matched sequences without sequence-level confirmation that each target anatomy is visible. Reports with multiple matched sequences contribute their labels multiple times under the sequence-weighted objective.

4 Deployment Attribution Framework

4.1 Three Complementary Analysis Axes

An observed released-to-local performance difference does not identify a single adaptation mechanism: it may reflect correction of systematic output shift, the value of local supervision, or a change in downstream decision reliability. Let b_{st} denote the released prediction for measurement t on local sequence s, and y_{st} its observed numeric reference. We therefore ask three layer-specific deployment questions: at the output layer, what share of a specified comparator gap is calibration-recoverable; at the learning layer, whether local labels add value beyond released predictions and whether anchoring supplies supervision-dependent inductive bias; and at the decision layer, whether numeric prediction performance supports reliable downstream assertions at useful coverage. Controlled comparisons of calibration versus replacement heads, absolute versus residual regression across patient budgets, and interval-aware versus point-based assertions operationalize these questions. The axes are distinct, complementary, and non-additive: the first two assess continuous error; the third assesses false assertions jointly with coverage.

4.2 Shared Evaluation Substrate

Following the official PanEcho input convention, a normalized 16-frame, 224×224 clip is mapped by the frozen pretrained backbone to

$$\mathbf{h}_s = g_{\mathrm{PE}}(\mathbf{x}_s) \in \mathbb{R}^d. \tag{1}$$

Each target uses an independent compact regression head (LayerNorm–Linear$_d$–GELU–Dropout–Linear$_1$),

$$\hat{z}_{st} = h_t(\mathbf{h}_s). \tag{2}$$

Training-only μ_t and σ_t define $z_{st} = (y_{st} - \mu_t)/\sigma_t$ and restore clinical units through $\hat{y}_{st} = \sigma_t \hat{z}_{st} + \mu_t$.

With missing-label mask m_{st}, each task uses SmoothL1,

$$\ell_t = \frac{1}{\sum_{s \in \mathcal{B}} m_{st}} \sum_{s \in \mathcal{B}} m_{st}\ \mathrm{SmoothL1}(\hat{z}_{st}, z_{st}), \tag{3}$$

skipping tasks without a valid batch label and combining losses after per-task averaging with inverse-frequency weights. Thus, "Joint" is a shared training container with independent trainable heads while the representation stays frozen. For reports with multiple sequences, the report-level estimand averages all sequence predictions for each target.

4.3 Calibration Attribution

Without changing the frozen representation, low-complexity affine calibration ($\hat{y}' = a\hat{y} + b$) provides an adaptation baseline for target-wise output shift; intercept calibration uses $\hat{y}' = \hat{y} + b$, and fixed-regularization ridge uses the same frozen features. All methods share finite test rows, and nMAE uses training-set standard deviations.

Let R_{direct}, R_{affine}, and R_{local} denote macro nMAE for the released predictions, train-only affine calibration, and the historical primary local MLP, respectively. Define the specified released-to-local gap and its calibration-recoverable portion as

$$\Delta_{\mathrm{local}} = R_{\mathrm{direct}} - R_{\mathrm{local}}, \qquad \Delta_{\mathrm{cal}} = R_{\mathrm{direct}} - R_{\mathrm{affine}}. \tag{4}$$

Their ratio is

$$\rho_{\mathrm{cal}} = \frac{\Delta_{\mathrm{cal}}}{\Delta_{\mathrm{local}}} = \frac{R_{\mathrm{direct}} - R_{\mathrm{affine}}}{R_{\mathrm{direct}} - R_{\mathrm{local}}}. \tag{5}$$

Thus, ρ_{cal} is the calibration-recoverable share relative to this specified comparator. Patient-cluster bootstrap recomputes target-wise nMAE, its macro average, and the ratio.

4.4 Supervision-Dependent Value of Released Predictions

The second analysis axis examines the supervision-dependent value of released predictions. Absolute regression relearns y_{st} from the frozen feature, whereas source-anchored regression learns its deviation from b_{st},

$$\delta_{st} = y_{st} - b_{st}, \qquad \hat{y}^{\mathrm{res}}_{st} = b_{st} + d_t(\mathbf{h}_s, b_{st}). \tag{6}$$

Residual scaling uses only the actual training-patient subset, validation selects the stopping epoch, and test labels enter final evaluation only. Linear, MLP, and MLP-with-explicit-b variants use raw or standardized SmoothL1. Because b_{st} is derived from $\mathbf{h}_s$, this comparison isolates the skip/anchoring parameterization under the same input evidence.

4.5 Decision-Level Uncertainty Propagation

The third analysis axis studies how predictive uncertainty affects downstream decision reliability.

Patient-group interval calibration. Because canonical validation selected C2 checkpoints, seed 42 splits canonical training into proper-training (4,477 patients; 5,566 rows) and calibration (1,119 patients; 1,431 rows); validation and test remain patient-disjoint and unchanged. The frozen C2 residual architecture is refit for 22 epochs using the validation-selected seed-42 schedule fixed before C3 test evaluation; full protocol details are in the appendix.

For target t and calibration patient p, patient-group nonconformity is the maximum absolute sequence residual,

$$s_{pt} = \max_{j \in p} |y_{ptj} - \hat{y}_{ptj}|. \tag{7}$$

For fixed $\alpha \in \{0.20, 0.10, 0.05\}$, $q_t(\alpha)$ is the finite-sample higher-rank quantile of these patient scores, and the sequence interval is

$$I_t(x) = [\hat{y}_t(x) - q_t(\alpha),\ \hat{y}_t(x) + q_t(\alpha)]. \tag{8}$$

The primary setting is $\alpha = 0.10$. We report sequence coverage and patient-all-sequence coverage. This target-wise split-conformal construction assumes patient exchangeability and remains fixed for downstream analysis.

Rule propagation and selective decisions. Let the prespecified rule for target t partition the measurement line into categories $\mathcal{C}_t$. The downstream object is the evidence-preserving chain

$$\mathbf{x}_s \longrightarrow \hat{y}_{st} \longrightarrow I_t(\mathbf{x}_s) \longrightarrow \mathcal{C}_t(I_t) \longrightarrow \{\texttt{assert}, \texttt{abstain}\}. \tag{9}$$

The compatible set $\mathcal{C}_t(I_t)$ contains every category intersected by the interval. We assert only when it has cardinality one; an interval crossing a threshold or spanning severity bins yields `uncertain`/`abstain`.

Primary analysis covers four provenance-verified single-measurement rules: LVEF, LVEDD, aortic sinus diameter, and aortic maximum velocity. The measurement-derived reference $r^{\mathrm{ref}} = f(y)$ applies each frozen rule to the observed measurement; multi-measurement interpretations remain outside scope.

The hard baseline emits $f(\hat{y})$. False assertion is error conditional on a definite output; assertion coverage is the fraction receiving one; and error capture is the fraction of hard-rule errors on which the interval-aware method abstains.

5 Empirical Investigation

5.1 Protocol

We use the patient-disjoint split in Section 3 and keep the PanEcho backbone frozen; heads use AdamW, SmoothL1, and early stopping. Normalized MAE (nMAE) divides target-wise MAE by training-set standard deviation, with macro averaging across targets. Direct and local-head comparisons use identical cached features and test rows for seven strict semantic/unit mappings. Confidence intervals use 10,000 patient-cluster bootstrap samples, and paired tests use per-patient mean absolute errors with Holm correction. The three axes evaluate calibration-recoverable gain, the supervision-dependent value of released predictions, and the false-assertion versus definite-coverage trade-off, with the last interpreted separately from continuous-measurement accuracy.

5.2 Local Retraining Reveals a Deployment Gap but Does Not Identify Its Source

Local heads lower sequence-level MAE for all seven strictly mapped tasks (Table 1); reductions range from 3.9% for aortic-root diameter to 25.8% for posterior-wall thickness, every bootstrap interval favors adaptation, and all paired tests remain significant after Holm correction. The unweighted mean reduction is 14.01% (patient-cluster bootstrap 95% CI, 12.59–15.37%). The largest gains occur for posterior-wall thickness and aortic-valve peak velocity; LVEF improves from 5.458 to 4.908 percentage points and LVIDd from 4.537 to 3.860 mm. These results establish a released-to-local gap while motivating calibration attribution, because replacement heads can absorb both global miscalibration and case-specific residual error.

5.3 Apparent Adaptation Gains Are Largely Explained by Calibration

Train-only affine calibration accounts for 74.5% of the macro normalized-MAE gap between released predictions and the historical primary local MLP (patient-cluster bootstrap 95% CI, 68.3–81.6%; 0.5503 direct, 0.4990 affine, and 0.4815 MLP; Table 2). The ridge-linear and intercept-only baselines

Table 1: Official zero-shot PanEcho heads versus newly trained local heads on the same frozen representation, test sequences, preprocessing, and strict task/unit mappings. MAE is sequence-weighted; 95% intervals use 10,000 patient-cluster bootstrap samples. Holm p values use paired per-patient mean absolute errors. Positive reduction favors local adaptation; pp denotes percentage points.

Target	Unit	N (seq./pat.)	Direct MAE	Local MAE	Rel. reduction [95% CI]	Holm p
Aortic root	mm	1495/1195	3.394	**3.263**	3.9 [2.1, 5.7]%	< 0.001
LVIDs	mm	1493/1193	4.152	**3.845**	7.4 [4.1, 10.5]%	< 0.001
LVPWd	mm	1497/1195	1.545	**1.146**	25.8 [21.9, 29.5]%	< 0.001
IVSd	mm	1496/1194	1.561	**1.382**	11.4 [7.8, 14.9]%	< 0.001
AV peak velocity	cm/s	1493/1193	43.042	**32.481**	24.5 [21.2, 27.8]%	< 0.001
LVIDd	mm	1485/1185	4.537	**3.860**	14.9 [11.5, 18.2]%	< 0.001
LVEF	pp	1501/1199	5.458	**4.908**	10.1 [7.2, 12.9]%	< 0.001

LVIDs/LVIDd: left-ventricular internal diameter at end systole/diastole; LVPWd: left-ventricular posterior-wall thickness at end diastole; IVSd: interventricular-septal thickness at end diastole; LVEF: left-ventricular ejection fraction; AV: aortic valve.

Table 2: Historical primary-checkpoint calibration decomposition over seven strict mappings. nMAE uses each target's training-set standard deviation; lower is better. Calibration and the linear head are fitted on training data only. The local MLP row is not the matched cached-feature seed-42 absolute retraining in Table 3.

Method	Macro nMAE
Official direct	0.5503
Official + intercept	0.5093
Official + affine	0.4990
Local linear	0.5104
Historical primary local MLP	0.4815

reach 0.5104 and 0.5093. Thus, systematic output correction explains a substantial portion of apparent adaptation improvement, making per-target affine calibration a lightweight diagnostic before case-specific nonlinear fitting.

The 74.5% estimate is specific to the macro normalized-MAE gap from released predictions to the historical primary local MLP. The remaining local-head advantage is heterogeneous across measurements: although the MLP has lower MAE point estimates than affine calibration for all seven targets, after Holm correction only aortic-valve peak velocity, interventricular-septal thickness, and LVEF retain support.

In the 28-field quality sensitivity analysis, a mask based on frozen training-only Tukey outer fences changes historical mean task-wise reduction from 14.0% (12.6–15.4%) to 15.0% (13.5–16.3%) and preserves direct $>$ affine $>$ MLP; affine calibration explains 81.1% of the robust gap (77.6–84.8%).

5.4 Released Predictions Provide Supervision-Dependent Value

On matched cached features, seed-42 macro nMAE is 0.4990/0.4713/0.4696 for affine/absolute/residual (Table 3). Residual minus absolute is -0.0016 (patient-cluster 95% CI, -0.0029 to -0.0004), only 0.35% of absolute nMAE; no target survives Holm correction, and explicit conditioning on the released output adds no meaningful improvement. Robust analysis retains the residual–affine gap, but its residual-minus-absolute interval crosses zero ($[-0.0039, 0.0004]$), characterizing the full-data difference as modest.

Under scarce supervision (Table 9 and Figure 5), 280 training patients (5%) give 0.4924 ± 0.0012, 0.5077 ± 0.0014, and 0.4959 macro nMAE for residual, absolute, and affine models. At this 5% budget, retaining the released prediction as a fixed output anchor provides additional value over matched absolute regression, consistent with a supervision-dependent inductive bias, while its value beyond affine calibration is small. The residual–absolute gap narrows from 0.0153 at 5% to roughly 0.0011 at 100%, where no target survives Holm correction.

Table 3: Matched cached-feature strict-seven sequence-level adaptation. Macro nMAE averages target-wise MAE normalized by training standard deviation. Absolute and residual rows are seed-42 retrainings on one execution path; they are distinct from the historical primary local MLP in Table 2. The residual-vs-absolute interval is a patient-cluster bootstrap.

Method	Macro nMAE	Evidence boundary
Official direct	0.5503	No local fitting
Official + affine	0.4990	Train-only global correction
Absolute MLP	0.4713	Matched seed 42
Residual MLP	**0.4696**	Matched seed 42

Residual minus absolute: -0.0016 (95% CI $[-0.0029, -0.0004]$); 0/7 target-wise comparisons pass Holm correction. Under the robust protocol the corresponding CI is $[-0.0039, 0.0004]$.

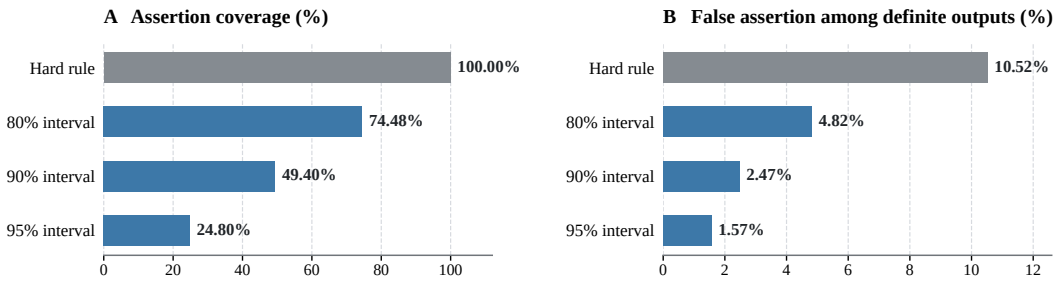


Figure 2: Risk–coverage trade-off against the measurement-derived rule reference. Panel A shows assertion coverage, and Panel B shows false assertion among definite outputs, for the always-assert point-rule baseline and uncertainty-aware rules with fixed 80%, 90%, and 95% interval settings. Abstention changes coverage, and the endpoint is selective consistency with categories derived from observed measurements.

5.5 Uncertainty Improves Downstream Decision Reliability

At the primary $\alpha = 0.10$ setting, empirical patient-all-sequence interval coverage is 90.99% for LVEF, 91.22% for LVEDD, 88.54% for aortic sinus diameter, and 89.94% for aortic maximum velocity. Aortic-sinus coverage is mildly below nominal.

At the primary setting, the uncertainty-aware rule has 2.47% false assertion among definite outputs at 49.40% assertion coverage, versus 10.52% at 100% coverage for the point rule, and captures 79.28% of hard-rule errors (Figure 2). Across fixed 80%, 90%, and 95% intervals, coverage decreases from 74.48% to 49.40% to 24.80%, while false assertion decreases from 4.82% to 2.47% to 1.57%. This rate applies only to definite outputs: abstention removes both hard-rule errors and correct cases. Continuous predictions remain unchanged, and diagnostic accuracy requires independent clinical labels.

The trade-off is strongly target-dependent. At the primary setting, LVEF yields only 5.00% definite assertion coverage, compared with 46.20% for LVEDD, 53.91% for aortic sinus diameter, and 92.50% for aortic velocity. The very low LVEF definite coverage is an important practical limitation: near clinical thresholds, the available measurement evidence often supports multiple rule categories.

6 Discussion and Conclusion

6.1 Deployment Implications

A zero-shot–adapted comparison establishes that local performance changed, but it does not identify why. We analyze deployment gains along three complementary dimensions: output calibration, the supervision-dependent value of released predictions, and decision-level uncertainty. This is a controlled attribution analysis rather than an additive identity: the first two concern continuous regression error, while the third changes the risk–coverage operating point of downstream decisions. Reporting only the endpoints therefore conflates distinct mechanisms and estimands.

Calibration before adaptation. A local head can combine systematic output correction with case-specific fitting. In this PanEcho cohort, train-only affine calibration recovers a substantial share of the specified gap without changing the frozen representation, separating output-shift correction from case-specific adaptation.

Released predictions as deployment anchors. The value of retaining released predictions is clearest with scarce local labels, while the full-data residual–absolute difference is small and has limited target-wise support. This value is supervision-dependent: a released prediction can act as a useful anchor without adding information beyond the frozen representation.

Decision-level uncertainty. Prediction error and decision reliability are distinct endpoints. Patient-group intervals reduce incorrect definite rule categories through abstention, requiring false assertion to be reported jointly with coverage and separately from numeric accuracy. This axis characterizes a deployment effect that a regression-only comparison cannot capture.

6.2 Resource-Aware Deployment

Affine calibration estimates only a per-target scale and intercept and may serve as a lightweight first diagnostic. Direct comparisons of GPU use, training time, and memory remain future work.

6.3 Limitations

External validity. The retrospective, single-center cohort weakly aligns report measurements to sequences under a sequence-weighted objective and evaluates one released model, PanEcho. This evidence characterizes institution-specific deployment of PanEcho; transportability across institutions, report schemas, acquisition protocols, and model families lies outside the present empirical scope.

Clinical validation. C3 covers four prespecified single-measurement rules whose references apply the same frozen rule to observed measurements. The endpoint is consistency with measurement-derived categories; diagnosis-level validity requires independent clinical outcomes. Patient-group split conformal calibration targets marginal coverage under exchangeability.

6.4 Ethics Statement

This study was approved by the Ethics Review Committee of Beijing Anzhen Hospital and was conducted in accordance with the Declaration of Helsinki and relevant institutional ethical guidelines. Given the retrospective design, the ethics committee waived the requirement for written informed consent. All echocardiographic examination data and related clinical information were anonymized before analysis.

6.5 Conclusion

We introduce a controlled deployment attribution framework that separates calibration-recoverable improvement, the supervision-dependent value of released predictions, and decision-level uncertainty effects. In this single-center PanEcho study, affine calibration recovers 74.5% of the macro normalized-MAE gap to the historical primary local MLP; released predictions provide their clearest additional value under scarce local supervision; and patient-group intervals expose a false-assertion versus definite-coverage trade-off under four fixed measurement-derived rules. This controlled analysis supports interpreting observed gains through output calibration, supervision-dependent value, and decision-level uncertainty before invoking more complex adaptation strategies.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Francesco Dalla Serra, William Clackett, Hamish MacKinnon, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q. O'Neil. Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations. In *Proceedings of the 2nd Conference of the*

Asia-Pacific Chapter of the Association for Computational Linguistics, pp. 615–624, 2022. doi: 10.18653/v1/2022.aacl-main.47.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. doi: 10.1038/s42256-021-00338-7. URL https://doi.org/10.1038/s42256-021-00338-7.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360, 2022. doi: 10.18653/v1/2022.findings-emnlp.319.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(S1):1513–1589, 2023. doi: 10.1007/s10462-023-10562-9. URL https://doi.org/10.1007/s10462-023-10562-9.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Gregory Holste, Evangelos K. Oikonomou, Márton Tokodi, Attila Kovács, Zhangyang Wang, and Rohan Khera. Complete AI-enabled echocardiography interpretation with multitask deep learning. *JAMA*, 2025. doi: 10.1001/jama.2025.8731. URL https://doi.org/10.1001/jama.2025.8731.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2132–2141, 2018.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018. doi: 10.1109/CVPR.2018.00781.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Roberto M. Lang, Luigi P. Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A. Flachskampf, Elyse Foster, Steven A. Goldstein, Tatiana Kuznetsova, Patrizio Lancellotti, Denisa Muraru, Michael H. Picard, Ernst R. Rietzschel, Lawrence Rudski, Kirk T. Spencer, Wendy Tsang, and Jens-Uwe Voigt. Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the american society of echocardiography and the european association of cardiovascular imaging. *Journal of the American Society of Echocardiography*, 28(1):1–39.e14, 2015. doi: 10.1016/j.echo.2014.10.003. URL https://doi.org/10.1016/j.echo.2014.10.003.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://doi.org/10.1080/01621459.2017.1307116.

Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *npj Digital Medicine*, 1(1), 2018. doi: 10.1038/s41746-017-0013-1. URL https://doi.org/10.1038/s41746-017-0013-1.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5288–5304, 2021. doi: 10.18653/v1/2021.naacl-main.416. URL https://aclanthology.org/2021.naacl-main.416/.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. doi: 10.1038/s41586-023-05881-4. URL https://doi.org/10.1038/s41586-023-05881-4.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. doi: 10.1038/s41586-020-2145-8.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/eb1e78328c46506b46a4ac4a1e378b91-Abstract.html.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5897–5906, 2019. URL https://proceedings.mlr.press/v97/song19a.html.

Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016. doi: 10.1109/TMI.2016.2535302. URL https://doi.org/10.1109/TMI.2016.2535302.

Milos Vukadinovic, I-Min Chiu, Xiu Tang, Neal Yuan, Tien-Yu Chen, Paul Cheng, Debiao Li, Susan Cheng, Bryan He, and David Ouyang. Comprehensive echocardiogram evaluation with view primed vision language AI. *Nature*, 650(8103):970–977, 2025. doi: 10.1038/s41586-025-09850-x. URL https://doi.org/10.1038/s41586-025-09850-x.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018. doi: 10.1371/journal.pmed.1002683. URL https://doi.org/10.1371/journal.pmed.1002683.

Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H. Tison, Laura A. Hallock, Lauren Beussink-Nelson, Mats H. Lassen, Eugene Fan, Mandar A. Aras, ChaRandle Jordan, Kirsten E. Fleischmann, Michelle Melisko, Atif Qasim, Sanjiv J. Shah, Ruzena Bajcsy, and Rahul C. Deo. Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy. *Circulation*, 138(16):1623–1635, 2018. doi: 10.1161/CIRCULATIONAHA.118.034338. URL https://doi.org/10.1161/CIRCULATIONAHA.118.034338.

AI USE STATEMENT

Generative-AI tools, including Codex, were used to assist code development, experiment execution and consistency checks, as well as language editing during manuscript preparation. The tools were used for implementation support, workflow assistance, and writing refinement; they did not determine the study design, generate scientific conclusions, or create patient data, clinical labels, reference outcomes, or experimental measurements. The authors reviewed and verified all code, artifacts, analyses, and text, and remain responsible for the study design, validity, interpretation, and final manuscript.

A Additional Results and Reproducibility

The repository-relative commands below assume that dependencies have been installed and the project Python environment has been activated. Environment names and machine-specific paths are intentionally omitted.

A.1 Distinct Recommended Targets

Table 4: Per-target results for the 16 distinct signals in the deduplicated recommended set. Training coverage is the number of finite training labels divided by 6,997 sequence rows. The configured 17-field run contained one exact duplicate: the field labeled pulmonary-artery peak velocity was identical to main pulmonary-artery diameter in all 10,059 jointly observed rows and is excluded here. pp denotes percentage points.

Target	Unit	Train N (coverage)	Test N	MAE	nMAE	R^2	Pearson
Aortic root	mm	6973 (99.7%)	1495	3.263	0.661	0.225	0.477
LVIDs	mm	6969 (99.6%)	1493	3.845	0.541	0.448	0.672
LVPWd	mm	6976 (99.7%)	1497	1.146	0.118	0.097	0.328
IVSd	mm	6965 (99.5%)	1496	1.382	0.586	0.216	0.481
RVOT diameter	mm	6962 (99.5%)	1495	2.806	0.586	0.142	0.390
AV peak velocity	cm/s	6928 (99.0%)	1493	32.481	0.449	0.250	0.502
RV AP diameter	mm	6949 (99.3%)	1493	2.494	0.649	0.135	0.379
Main PA diameter	mm	6928 (99.0%)	1484	2.377	0.504	0.132	0.367
LVIDd	mm	6960 (99.5%)	1485	3.860	0.563	0.426	0.657
LVEF	pp	6997 (100.0%)	1501	4.908	0.452	0.391	0.628
Mitral E velocity	cm/s	6897 (98.6%)	1488	20.618	0.195	0.328	0.582
Septal excursion	mm	6872 (98.2%)	1479	1.256	0.553	0.064	0.253
Ascending aorta	mm	6786 (97.0%)	1441	3.463	0.657	0.200	0.450
Mitral A velocity	cm/s	6188 (88.4%)	1345	18.285	0.706	0.139	0.382
Fractional shortening	pp	5433 (77.6%)	1139	2.637	0.648	0.163	0.413
Left atrial diameter	mm	3100 (44.3%)	671	3.443	0.645	0.302	0.561

The configured 17-field run contains one exact duplicate signal. Excluding it gives macro nMAE 0.5320 and mean R^2 0.2285 over 16 distinct recommended signals; all have positive R^2 and six have Pearson correlation above 0.5. Single-task, anatomical-group, and joint-head point estimates are close for 15 distinct core signals, but one seed and no non-inferiority margin preclude an equivalence claim. The field audit finds 28 candidates representing 27 distinct signals, 25 distinct trainable signals, and 20 broad test-evaluable signals. Core, recommended, and broad-evaluable macro nMAEs are 0.5224, 0.5320, and 0.5752 over different target mixtures, so they do not identify a causal effect of task breadth.

A.2 Task-Wise Relative Improvement

A.3 Target-Set Definitions

The core set requires at least 1,000 matched valid examples, no more than 40% missingness, an explicit unit, and clear medical meaning. The recommended set requires at least 500 matched valid examples and no more than 70% missingness. The broad candidate set uses at least 100 matched valid examples before the final manifest-column check. Missing targets are never filled with zero or a population mean, and a sample contributes to every task for which its label is finite.

A.4 Statistical Analysis

The paired analysis is reproduced by

```
python \
  scripts/bootstrap_panecho_comparison.py \
  --bootstrap-samples 10000 --seed 42
```

The script verifies one-to-one prediction-row alignment and equality of the target values before computing errors. Bootstrap resampling uses patient identifiers as clusters, while the reported MAE

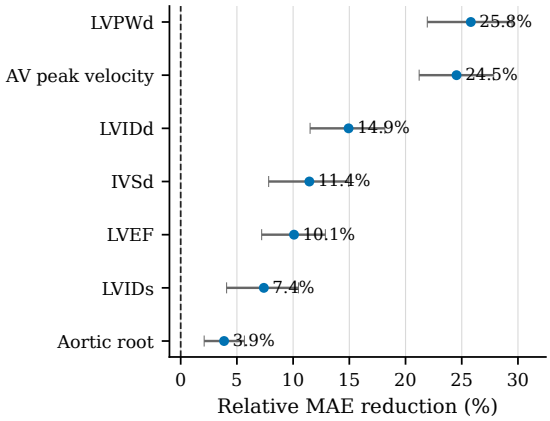


Figure 3: Task-wise relative reduction in sequence-level MAE for newly trained local regression heads versus official zero-shot PanEcho heads on the same patient-disjoint test split and frozen representation. Positive values favor local adaptation. Error bars are 95% patient-cluster bootstrap intervals over 10,000 resamples for seven strict task/unit mappings; all intervals exclude zero.

Table 5: Sequence- and report-level strict-seven estimands for released PanEcho and the historical primary local MLP. Report pooling gives each report one vote; it is not learned multi-view fusion.

Estimand	Released nMAE	Historical local nMAE
Sequence weighted	0.5503	0.4815
Report mean	0.5372	0.4703
Report median	0.5373	0.4710

remains sequence-weighted within a sampled cluster set. The Wilcoxon analysis first averages absolute error within each patient.

Calibration decomposition and report-level estimands are reproduced by

```
python \
  scripts/iclr_revision/evaluate_calibration_baselines.py \
  --bootstrap-samples 10000 --seed 42
python \
  scripts/iclr_revision/evaluate_report_level_primary.py \
  --bootstrap-samples 10000 --seed 42
```

Both scripts assert official/local row and target equality. Calibration parameters are fitted on training rows only; normalized MAE uses the corresponding training-target standard deviation. Report bootstrap samples patients and retains all of each sampled patient's reports.

A.5 C3 Uncertainty Protocol and Per-Rule Results

The rule inventory is read from the frozen deterministic implementation and records its source hash. Primary C3 uses only LVEF (thresholds 40/50/55%), LVEDD (55/60 mm), aortic sinus diameter (40 mm), and aortic maximum velocity (250/400 cm/s). These are the four strict-seven targets with provenance-verified single-measurement rules. These thresholds are frozen operational interpretation rules inherited from the existing local deterministic implementation; source-hash verification establishes reproducibility of the rule set, not independent clinical adjudication of the thresholds. LVESD consistency and wall-thickening depend on multiple measurements and are excluded because exact joint interval propagation has not been validated; no threshold was added or modified.

Canonical validation selected the original checkpoints, so it is not reused for calibration. The fixed seed-42 split of canonical training contains 4,477 proper-training patients (5,566 rows) and 1,119

calibration patients (1,431 rows). Canonical validation and test remain untouched, and all four partitions are patient-disjoint. The C3 residual predictor reuses the frozen 768–256 C2 architecture and optimization settings for 22 fixed epochs on proper-training data; this count is inherited from the validation-selected epoch of the frozen C2 seed-42 residual checkpoint and was fixed before C3 test evaluation. Training uses neither early stopping nor the backbone. Calibration then uses one maximum absolute residual score per patient and target. The primary $\alpha = 0.10$ and sensitivity $\alpha = 0.20/0.05$ values were fixed before test evaluation.

Table 6: Patient-aware interval calibration for the four primary rules. Calibration uses 1,119 patients per target. Empirical coverage is the fraction of test patients for whom all labeled sequences fall within the interval; width is in the displayed unit.

Target	Nominal	Test rows	Test patients	Empirical cov.	Mean width
LVEF (%)	80%	1,501	1,199	81.65%	15.40
	90%			90.99%	22.39
	95%			95.00%	30.14
LVEDD (mm)	80%	1,485	1,185	81.27%	12.42
	90%			91.22%	16.72
	95%			95.53%	20.80
Aortic sinus (mm)	80%	1,495	1,195	78.08%	9.67
	90%			88.54%	12.83
	95%			93.72%	15.65
Aortic velocity (cm/s)	80%	1,493	1,193	78.12%	86.19
	90%			89.94%	140.38
	95%			96.23%	233.04

Table 7: Finding-level risk–coverage sensitivity. False assertion is computed among definite outputs; error capture is the fraction of always-assert hard-rule errors on which the uncertainty-aware rule abstains. Hard false-assertion rates are 15.66%, 11.11%, 10.03%, and 5.29% for LVEF, LVEDD, aortic sinus, and aortic velocity.

Target	Interval	Assertion cov.	Abstention	False assertion	Error capture
LVEF	80%	49.50%	50.50%	4.85%	84.68%
	90%	5.00%	95.00%	0.00%	100.00%
	95%	0.00%	100.00%	–	100.00%
LVEDD	80%	72.26%	27.74%	4.57%	70.30%
	90%	46.20%	53.80%	2.92%	87.88%
	95%	21.08%	78.92%	0.64%	98.79%
Aortic sinus	80%	79.80%	20.20%	6.12%	51.33%
	90%	53.91%	46.09%	4.22%	77.33%
	95%	28.03%	71.97%	2.86%	92.00%
Aortic velocity	80%	96.38%	3.62%	3.75%	31.65%
	90%	92.50%	7.50%	2.75%	51.90%
	95%	50.10%	49.90%	1.20%	88.61%

At 90%, macro unnecessary abstention is 47.81%: this is the fraction of hard-rule-correct rows on which the interval-based rule abstains. Accordingly, error capture is reported jointly with coverage and unnecessary abstention rather than treating all abstentions as correct. Near-boundary is defined without test truth as point-estimate distance to the nearest threshold divided by the calibration radius being at most one. Near-boundary hard-rule error rates are 16.48% for LVEF, 18.15% for LVEDD, 16.84% for aortic sinus, and 36.61% for aortic velocity at the primary radius. Wide intervals, rather than a hidden change in rule logic, drive the low LVEF assertion coverage.

The portable evaluation path is

```
python \
 scripts/iclr_revision/c3_uncertainty/\
 calibrate_prediction_intervals.py
python \
 scripts/iclr_revision/c3_uncertainty/\
 propagate_intervals_to_findings.py
python \
 scripts/iclr_revision/c3_uncertainty/\
 evaluate_uncertainty_findings.py
```

```
python \
 scripts/iclr_revision/c3_uncertainty/\
 analyze_risk_coverage.py
```

The split manifest, fixed quantiles, intervals, rule states, finding metrics, and failure attribution are retained as versioned artifacts. No private absolute path is required.

A.6 Target-Independent Split Audit

The target-independent patient split is generated over all matched rows by

```
python \
  scripts/iclr_revision/build_target_independent_split.py \
  --seed 42
```

It asserts disjoint patient, report, study, and composite-sequence identifiers. A newly extracted isolated cache covers all 10,175 rows and records source-manifest and PanEcho-weight hashes. The strict-seven one-seed sensitivity is reproduced by

```
CACHE=outputs/iclr_revision/target_independent_feature_cache
RUN=outputs/iclr_revision/target_independent_retraining
PANECHO_WEIGHT=/path/to/author-supplied/panecho_checkpoint.pt
python scripts/extract_features.py \
  --targets metadata/targets_panecho_strict7.json \
  --split-dir metadata/iclr_revision/\
target_independent_split --split-suffix="" \
  --output-dir "$CACHE" --panecho-weight "$PANECHO_WEIGHT" \
  --batch-size 8 --num-workers 4 --seed 42
python \
  scripts/iclr_revision/train_feature_replication.py \
  --cache-dir "$CACHE" \
  --targets metadata/targets_panecho_strict7.json \
  --output-dir "$RUN/strict7_mlp_inverse_seed42" \
  --seed 42 --subset-seed 20260803 --architecture mlp \
  --task-weighting inverse_frequency
python \
  scripts/iclr_revision/evaluate_calibration_baselines.py \
  --cache-dir "$CACHE" \
  --manifest metadata/iclr_revision/\
target_independent_split/test.csv \
  --mlp-predictions "$RUN/strict7_mlp_inverse_seed42/\
predictions.csv" \
  --output-dir "$RUN/calibration"
```

A.7 Optional Renderer Containment Audit

The validator suite and renderer inputs are reproduced without an API call by

```
python \
  scripts/iclr_revision/build_contract_stress_suite.py
python \
  scripts/iclr_revision/evaluate_contract_stress_suite.py
python \
  scripts/iclr_revision/sample_llm_evaluation_cases.py --seed 42
RUN=scripts/iclr_revision
OUT=outputs/iclr_revision/renderer_ablation/full_200
env -u http_proxy -u https_proxy -u HTTP_PROXY \
  -u HTTPS_PROXY -u ALL_PROXY \
  python "$RUN/run_renderer_ablation.py" \
```

Table 8: Random-patient renderer component ablation on 200 patients. Rates are case proportions. Validator acceptance measures contract containment; it is not an autonomous diagnostic-accuracy metric.

Condition	Parse	Accept	Exact	Unsup. code	Unsup. num.	Latency (s)
Template	1.000	1.000	0.995	0.000	0.000	0.0
Measurement-only LLM	1.000	0.000	0.000	1.000	0.035	19.4
Typed-contract LLM	1.000	0.540	0.000	0.455	0.005	17.9
Contract + rule definitions	1.000	0.925	0.970	0.025	0.065	14.3
Contract + deterministic findings	1.000	1.000	0.995	0.000	0.000	9.5
Findings + validator	1.000	1.000	0.995	0.000	0.000	9.5
Findings + structured output	1.000	1.000	0.995	0.000	0.000	9.4

```
  --model-alias submission_generic --execute --resume \
  --output "$OUT/records.jsonl"
SUMMARY="$RUN/summarize_renderer_ablation.py"
python "$SUMMARY" \
  --records "$OUT/records.jsonl" \
  --bootstrap-samples 10000 --seed 42
```

The formal experiment uses one randomly selected case per patient and sends only direct-identifier-excluded prediction-only contracts. Reference measurements remain local for evaluation and are removed by an asserted serialization boundary. Generated records include local patient linkage, prompt hashes, decoding settings, raw provider responses, token use, parser status, validator output, attempts, latency, and timestamps. Credentials are inherited from environment variables and are never serialized into the formal records. The seven-condition run contains 1,400 complete records; all 1,200 live calls report the provider model as GPT-5.5 and parse successfully.

The portable command reads `LLM_API_KEY`, `LLM_BASE_URL`, and `LLM_MODEL` from the caller's environment. The sanitized submission package must exclude credentials, endpoint URLs, provider-specific secret mappings, raw error records containing URLs, and private linkage files; the full private research tree is not a distributable artifact.

A.8 Target-Quality Sensitivity Commands

The target audit, frozen statistical protocol, isolated robust cache, matched retraining, and decomposition are reproduced by

```
python \
  scripts/iclr_revision/audit_target_outliers.py
python \
  scripts/iclr_revision/apply_target_cleaning_protocol.py
python \
  scripts/iclr_revision/train_feature_replication.py \
  --cache-dir outputs/iclr_revision/cleaning_sensitivity/\
robust_feature_cache \
  --targets metadata/targets_panecho_strict7.json \
  --output-dir outputs/iclr_revision/cleaning_sensitivity/\
robust_strict7_mlp_seed42 --seed 42 --architecture mlp \
  --task-weighting inverse_frequency
python \
  scripts/iclr_revision/evaluate_cleaning_sensitivity.py \
  --bootstrap-samples 10000 --seed 42
```

The protocol explicitly marks clinical-range filtering as pending verified clinical input. The reported robust analysis is statistical sensitivity only.

A.9 Auxiliary Seed Sensitivity

Three matched cached-feature strict-seven MLP runs with seeds 42, 43, and 44 improve all seven targets in every run; their unweighted mean task-wise MAE reductions are 15.87%, 15.99%, and

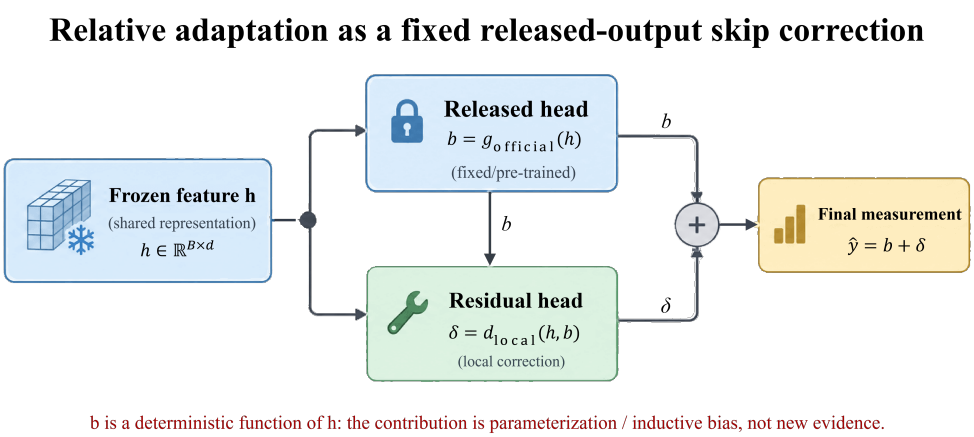


Figure 4: Residual local adaptation keeps the released prediction as a fixed skip term and learns a sample-specific correction. Because the released output is a deterministic function of the same frozen feature, the comparison isolates parameterization and optimization bias under the same input information.

Table 9: Macro nMAE under matched patient-level label fractions (mean ± standard deviation over seeds 42–44). Seed variation is optimization variability, not a patient confidence interval; bold marks the lowest macro point estimate and does not imply target-wise significance.

Labels	Affine	Absolute MLP	Residual MLP
5%	0.4959	0.5077±0.0014	**0.4924±0.0012**
10%	0.5009	0.4928±0.0008	**0.4847±0.0004**
25%	0.5005	0.4836±0.0003	**0.4789±0.0004**
50%	0.5010	0.4760±0.0004	**0.4744±0.0010**
100%	0.4990	0.4706±0.0006	**0.4696±0.0002**

16.06%. Equal and inverse-frequency weighting give effectively identical results because the representation is frozen and heads are independently parameterized. The runs use a separate feature-head execution path and therefore are not image-decoding replications of the historical primary checkpoint or pooled with its patient-cluster confidence interval.

A.10 Label-Efficiency Commands

Residual full-data, robust, matched inference, and label-efficiency artifacts are reproduced by

```
python \
 scripts/iclr_revision/relative_regression/train_residual_head.py \
 --output-dir outputs/iclr_revision/relative_regression/\
strict7_residual_mlp_seed42 --architecture mlp \
 --residual-scaling standardized --seed 42
python \
 scripts/iclr_revision/relative_regression/evaluate_relative.py \
 --bootstrap-samples 10000 --seed 42
python \
 scripts/iclr_revision/relative_regression/run_label_grid.py \
 --fractions 0.05,0.10,0.25,0.50,1.00 \
 --seeds 42,43,44 --architectures linear,mlp
python \
 scripts/iclr_revision/relative_regression/evaluate_label_grid.py \
 --bootstrap-samples 2000
```

Every residual scaler is fitted on the actual training-patient subset. Stored patient-list hashes are asserted against matched absolute runs. The patient-cluster intervals condition on a trained seed; seed standard deviations measure optimization variability and are not confidence intervals.

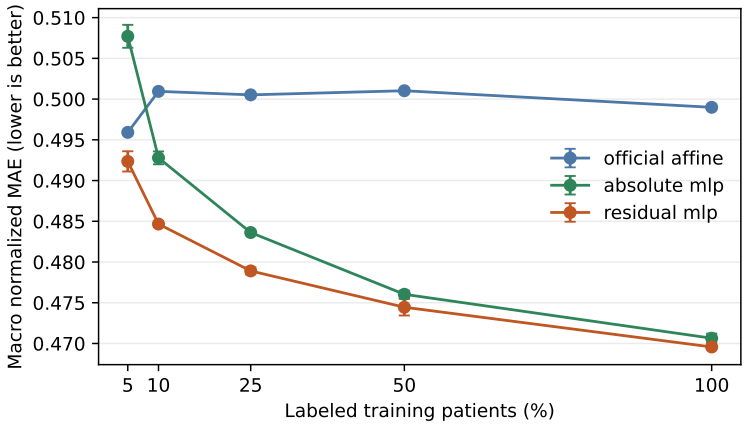


Figure 5: Strict-seven label efficiency under fixed patient-level subsets. Lines show macro nMAE and error bars show one standard deviation over three optimization seeds. Released-output anchoring provides its clearest value relative to absolute prediction in the lowest-label regimes; differences narrow with more local labels.

The minimum label-efficiency grid is reproduced by

```
python \
  scripts/iclr_revision/run_label_efficiency.py \
  --fractions 0.05,0.10,0.25,0.50,1.0 --seeds 42,43,44 \
  --architectures linear,mlp --subset-seed 20260803
python \
  scripts/iclr_revision/summarize_label_efficiency.py
```

Sampling is by unique patient, and every method at a fraction uses the same stored patient-list hash. Official intercept and affine parameters use only those training rows. Validation and test sets remain unchanged.

This cached-feature grid is a matched implementation check for the residual comparison; the main label-efficiency figure reports the claim-bearing residual, absolute, and affine results. Its 100% local endpoint is not the historical primary checkpoint.

Corrected simplest-sufficient selector. A validation-only selector over released, affine, and residual predictions uses the prespecified 1% validation tolerance and chooses the simplest method within tolerance of the validation winner. It selects released for 0/7 targets, affine for 0/7, and residual for 7/7; selected test macro nMAE therefore equals always-residual (approximately 0.4696). This negative result provides no complexity reduction and is not a main contribution.

A.11 Source-DICOM Execution and Orchestration Interfaces

An earlier manifest smoke writes stage hashes and statuses from aligned precomputed measurements; the audit below instead executes image inference from source DICOM. The collection contains 4,160 Part-10 US objects in 224 Studies and 225 Series: 1,348 Ultrasound Multi-frame Image Storage cine objects and 2,812 single-frame Ultrasound Image Storage objects, with no Secondary Capture. JPEG Baseline and Explicit VR Little Endian account for 3,190 and 970 objects. It is not established as the source of the full JPEG cohort or as patient-independent, and provides no outcome truth.

Eligibility is determined before model execution solely by the prespecified Ultrasound Multi-frame Image Storage SOP class; decode or inference success is not part of the denominator. All 1,348 eligible cine objects complete decode, EchoPrime predicted-view inference, PanEcho released-head inference, and frozen residual inference. The existing uniform sampler deterministically pads 120 cines below 16 frames; 561 have 16–63 frames and 667 have at least 64. Ultrasound-region cropping applies to 1,339 cines, with 9 metadata-unavailable fallbacks. The 80 remaining Studies terminate as `NO_MULTIFRAME_CINE`; they may contain single-frame US objects but no eligible multi-frame cine. No eligible Study has a decode, model, contract, rule, validator, or local-renderer failure. Median/p95 runtime is 16.93/62.54 seconds among local successes on CPU. Every Study retains

Table 10: Source-DICOM technical execution. Eligibility is assigned before decoding or inference by SOP class. Cine-level execution is aggregated to Study success; post-eligibility and overall denominators are 144 and 224. The endpoint uses execution status in an outcome-unlabeled cohort.

Stage	Eligible	Success	Coverage
Source Study record	224	224	100.0%
Eligible multi-frame cine present	224	144	64.3%
Cine decode (Study success)	144	144	100.0%
EchoPrime predicted view (Study)	144	144	100.0%
PanEcho released measurement (Study)	144	144	100.0%
Frozen residual adaptation (Study)	144	144	100.0%
Typed contract and rules	144	144	100.0%
Validator and local renderer	144	144	100.0%
Local structured workflow	224	144	64.3%

the frozen all-sequence routing fallback, and external rendering is recorded as `WITHHELD_GOVERNANCE`. No truth is loaded and no external LLM is called.

Public workflow artifacts use keyed study/object hashes; a validator finds zero configured identifier, raw-UID, private-path, or private-hash violations in 152 JSON/JSONL/CSV files. However, the private 100-frame pixel-PHI package is not manually adjudicated, and metadata include 589 `BurnedInAnnotation=YES` objects plus 1,541 missing/unknown values. Pixel-level de-identification therefore remains unverified, and no external renderer receives DICOM-derived content.

The full execution is reproduced by

```
python \
 scripts/iclr_revision/workflow/run_dicom_workflow.py \
 --panecho-weight /path/to/panecho_checkpoint.pt \
 --output-dir outputs/iclr_revision/dicom_workflow/full224 \
 --max-studies 0 --device cpu --resume
python \
 scripts/iclr_revision/workflow/summarize_dicom_workflow.py \
 --run-dir outputs/iclr_revision/dicom_workflow/full224
```

The PanEcho, residual, EchoPrime, routing, rule, and validator SHA256 values are recorded in `run_provenance.json`; source-pixel-derived model-input and feature hashes are retained per cine. The default or missing-checkpoint plane path in the older manifest orchestrator returns `unknown`; the voting fixture remains a software-conformance test, and the separate real model-version evaluation follows.

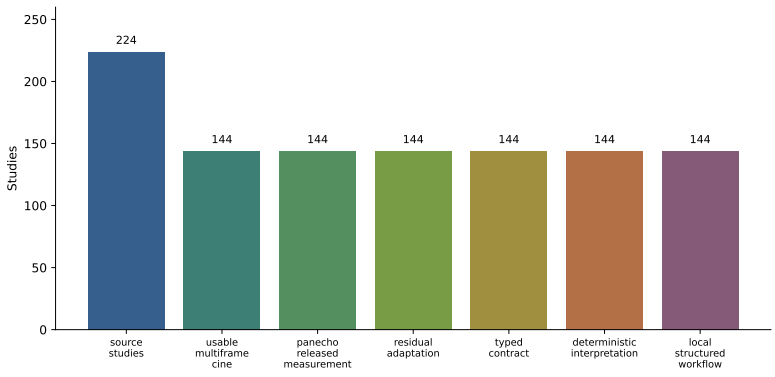


Figure 6: Stage-wise source-DICOM workflow execution. The 80 Studies outside the local endpoint contain no eligible multi-frame cine; all 144 eligible Studies complete the downstream local structured workflow. The endpoint is technical execution coverage over an outcome-unlabeled cohort.

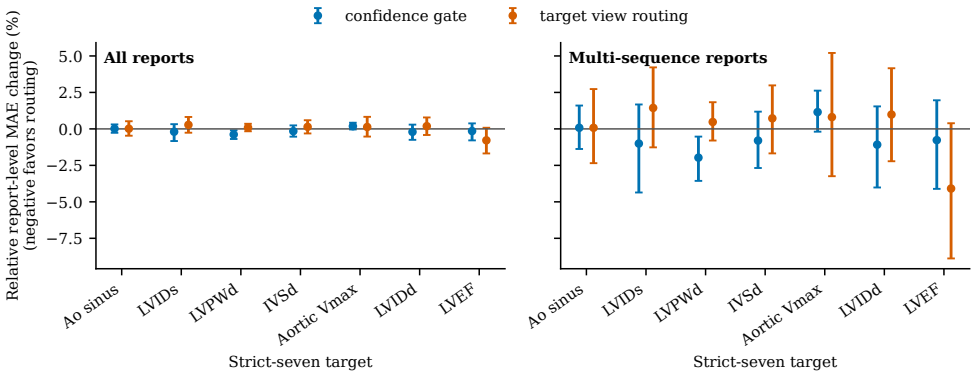


Figure 7: Relative held-out report-level MAE change, $100(\mathrm{MAE}_{route} - \mathrm{MAE}_{all})/\mathrm{MAE}_{all}$, for frozen EchoPrime-predicted-view routing versus all-sequence averaging. Negative values favor routing. Error bars transform the 10,000-sample patient-cluster bootstrap interval for the raw paired MAE difference by the observed all-sequence MAE; the denominator is not recomputed per bootstrap draw. Policies use training/validation data only, and no target survives Holm correction.

A.12 EchoPrime Predicted-View Reproducibility

The view classifier is taken from EchoPrime (Vukadinovic et al., 2025) commit `03874a5f203695f38968068c21584656d475b6b1`; the released checkpoint SHA256 is `046b76685de5295d151f74725d69be758c53aac3a028132e3f1f160ba5143954`. The official first-sampled-frame inference, cohort audit, and leakage-safe routing are reproduced by

```
MANIFEST=metadata/iclr_revision/\
echoprime_matched_manifest_with_split.csv
CKPT=third_party/EchoPrime/model_data/weights/view_classifier.pt
OUT=outputs/iclr_revision/plane_recognition/\
full_official_first_frame
python \
 scripts/iclr_revision/workflow/run_echoprime_view_inference.py \
 --manifest "$MANIFEST" --checkpoint "$CKPT" \
 --mode official_first_frame --output-dir "$OUT"
python \
 scripts/iclr_revision/workflow/analyze_echoprime_views.py
ROUTE=scripts/iclr_revision/workflow/\
evaluate_predicted_view_routing.py
VIEW=outputs/iclr_revision/plane_recognition/full_predictions.csv
python \
 "$ROUTE" --train-view-predictions "$VIEW" \
 --val-view-predictions "$VIEW" \
 --test-view-predictions "$VIEW" \
 --output-dir outputs/iclr_revision/plane_recognition/\
predicted_view_routing --bootstrap-samples 10000 \
 --min-train-patients 30 --seed 42
```

The complete classifier succeeds on 10,175/10,175 sequences. Of 8,283 reports, 1,467 contain multiple matched sequences. For the residual MLP on all reports, macro nMAE is 0.457788 for all-sequence averaging, 0.457348 after confidence gating, and 0.457867 after target-view routing; no target-wise improvement survives Holm correction. The frozen residual-model confidence gate covers 93.38% of target-report estimands overall; target-specific routing covers 24.03% and otherwise falls back to all-sequence averaging. On the multi-sequence subset these coverages are 98.14% and 36.63%. These are routing coverages, not anatomical-view correctness.

EchoPrime is distributed under the Cedars-Sinai Academic Software License rather than MIT. As required by its Paragraph 4: The Software used in this research was created by Milos Vukadinovic,

Table 11: Three same-gateway model-version outputs compared with a Level-D measurement-rule reference that shares the deterministic rule definition on 200 fixed random patients. Coverage is the fraction of all 200 cases receiving a prediction; abstained cases remain in its denominator. Mean F1 and exact agreement are unconditional except where marked as conditioned on covered cases.

Method	Mean case F1	Exact	Coverage	Validator accepted
Model A	1.000	1.000	1.000	1.000
Model B	0.999	0.995	1.000	1.000
Small model	0.690	0.425	1.000	0.785
Raw majority	0.999	0.995	1.000	–
Validated majority	0.999	0.995	1.000	–
Per-code 3/3 vote	1.000	1.000	1.000	–
Case disagreement abstention	1.000†	1.000†	0.425	–

Per-code 3/3 vote retains only codes with all three valid votes; it does not require identical full-case outputs and therefore has full case coverage here. †Conditioned on the 85 covered cases after strict case-level disagreement abstention. All 600 calls succeeded; versions share one gateway.

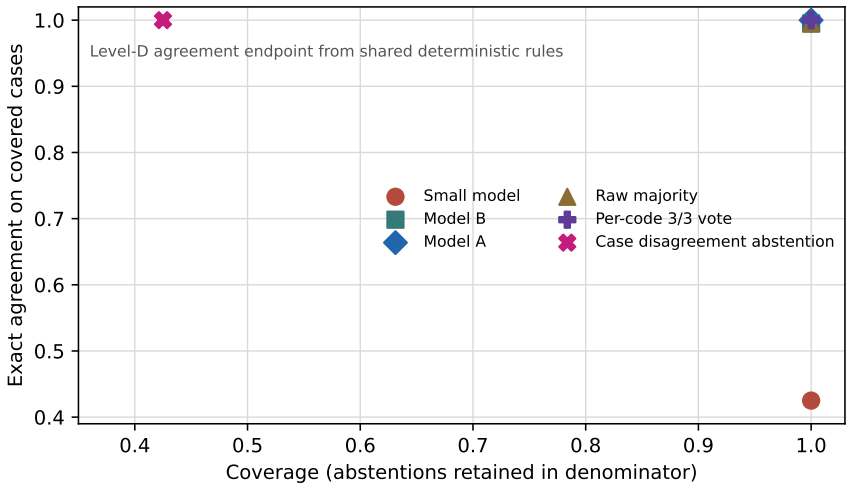


Figure 8: Coverage versus exact agreement with the Level-D measurement-rule reference, which shares the deterministic rule definition. Per-code 3/3 voting filters codes without requiring identical complete case outputs; strict case-level disagreement abstention retains 42.5% of cases. The strongest single model already reaches the reference ceiling. The curve characterizes agreement with this rule reference; clinical risk–coverage requires independent clinical labels.

Bryan He, and David Ouyang of Cedars-Sinai Medical Center. © 2024 Cedars-Sinai Medical Center. Paragraph 2 prohibits further transfer; the submission/public artifact therefore must not redistribute the EchoPrime source, archive, or checkpoint. Reproduction uses the official repository and release URL, with the commit and SHA256 above for verification.

A.13 Three-Model-Version Rule-Agreement Evaluation

A diagnosis-reference audit found no Level-A clinician-adjudicated or Level-B structured clinical diagnosis. The original free-text impression field can be converted into Level-C report-derived phenotypes only through a fixed regex pipeline; those labels have not completed manual audit and are excluded from clinical evaluation. The experiment below uses Level-D measurement-rule findings and therefore assesses structural agreement only.

The 600 calls comprise three independently executed provider-returned model versions on 200 fixed random patients. All calls succeed and parse. They use identical direct-identifier-excluded prediction-only contracts and cannot see reference report values, linkage identifiers, or other model outputs. Total usage is 781,368 tokens; mean and median per-call latency are 7.06 and 6.17 seconds, API

failure rate is zero, and retry rate is 0.17%. The versions share one gateway, and the four rule codes are highly imbalanced (one code has one positive case), preventing cross-provider or broad diagnostic claims.

A.14 Pipeline Conformance Test

For 100 cases, we replace only LVEF with a value on the opposite side of a rule threshold, recompute the deterministic rules, and pass the modified rule output to the renderer. Severity adherence and directional response are 100%. Because the expected rule is supplied to the renderer, this is a software conformance test rather than evidence of language-model reasoning or clinical robustness.

A.15 Additional Deployment Audits

Overlay masking. We evaluate the same trained heads under fixed masks covering left-top text, right-side scales, and the bottom text strip. Across the seven primary targets, mean relative MAE changes by -0.07%, and no target's unadjusted patient-cluster interval indicates degradation. An equal-area, path-seeded random mask worsens MAE by 9.87% on average; all seven unadjusted intervals exclude zero. Because these comparisons were not multiplicity-corrected, they are not confirmatory significance tests. The tested edge overlays are not necessary for performance, while removing generic image content is harmful; this does not exclude text or calipers outside the predefined regions.

Contract-validator stress suite. The deterministic validator makes the expected accept/reject decision on all 16 authored synthetic cases spanning threshold boundaries, missingness and status inconsistency, unit and magnitude errors, duplicate or conflicting evidence, missing provenance, and adversarial unsupported claims. False acceptance and false rejection are both zero. Since the validator itself blocks the injected errors, this measures software effectiveness on a small authored suite, not language-model robustness.

Last-block fine-tuning. We unfreeze PanEcho's final temporal transformer block, use a 10^{-5} backbone learning rate, and retain the same 15 deduplicated core targets. Relative to frozen training, the unweighted mean task-wise MAE improvement is 0.23% (patient-cluster 95% CI, -0.23% to 0.70%). Only aortic peak velocity remains significant after Holm correction, and it worsens by 5.29%. We therefore retain the frozen backbone.

A.16 Disease-Stratified Analysis Protocol

The versioned protocol derives overlapping report-text phenotypes, pools predictions within reports, and specifies patient-cluster inference, circularity flags, and minimum cell sizes. These labels are not expert adjudications; because their sampled manual audit remains incomplete, all subgroup estimates are deliberately excluded from the reported evidence.