CODELANCET: INTERNALIZING PHYSICIAN SKILLS FOR DIAGNOSIS VIA AGENTIC CODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) now approach specialists at recognizing what a medical image shows. Yet many diagnoses must be examined rather than recognized. On these, their correctness is often hollow. On ejection fraction, which must be measured by tracing the left ventricle through the cardiac cycle, the most accurate frontier VLMs miss 74–89% of failing hearts and score well only because most hearts are normal. What they lack is less medical knowledge than a means of acting on the image, so they fall back on the population prior. **Can a VLM learn to examine the way a physician does?** Our key insight is that executable code turns examination into a sequence of executed and inspectable acts on the image. Unlike a fixed toolset, code lets the model check and correct its own evidence, compose operations no designer anticipated, adapt its effort to the case and make perception itself an act, choosing where and how finely to look. We present CodeLancet, a 4B agent that examines medical images through executable code. Skill-guided, rubric-audited demonstrations and outcome-based GRPO first teach it to conduct the examination. Yet a trajectory-level outcome reward credits every step of a correct examination alike, and over a third of these contain critical wrong steps. Focal-PPO instead credits each decision by the change in expected outcome it causes, estimated by branching rollouts around the decisions where failed examinations first went wrong, and the second stage cuts this share to a tenth. Across six benchmarks in fundus photography, echocardiography and gigapixel pathology, CodeLancet outperforms six frontier VLMs on every headline metric: at comparable overall accuracy it detects 72.5% of failing hearts, it nearly halves referral errors in retinopathy, and on whole-slide melanoma images it issues about a quarter fewer false relapse alerts, with a margin that widens as the answer depends more on examination than on recognition. Blinded specialists endorse 80–86% of its steps. Our results suggest that diagnosis is learned not only by seeing better but also by acting on what is seen, and that code can be to a model what the lancet is to a physician.

1 INTRODUCTION

A physician does not read a medical image in a single glance. A cardiologist estimating ejection fraction traces the endocardium, identifies end-diastole and end-systole, and computes the fraction from chamber volumes (Lang et al., 2015); an ophthalmologist staging diabetic retinopathy surveys the four quadrants and counts lesions against the 4-2-1 rule (Wilkinson et al., 2003); a pathologist surveys a gigapixel slide and measures Breslow thickness and ulceration, the factors on which melanoma staging rests (Gershenwald et al., 2017). In each case, diagnosis is an *examination* rather than an act of recognition: each step acts on the image and yields evidence for the diagnosis.

Medical image models now match specialists at recognition (Gulshan et al., 2016; Esteva et al., 2017), and generalist vision–language models (VLMs) aim to extend it to open clinical questions (Moor et al., 2023). However, where a diagnosis depends on the examination itself, frontier VLMs show what we term *hollow correctness*: correctness that reflects the population rather than the patient. Ejection fraction makes this plain: it is a ratio of chamber volumes, measured by tracing the left ventricle at end-diastole and end-systole, not recognized at a glance. The three most accurate frontier VLMs are right on 66–70% of cases, yet detect only 11–26% of hearts with reduced ejection fraction ($\leq$40%); their accuracy is earned on the normal hearts that make up 78% of the test set (Table 8).

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.

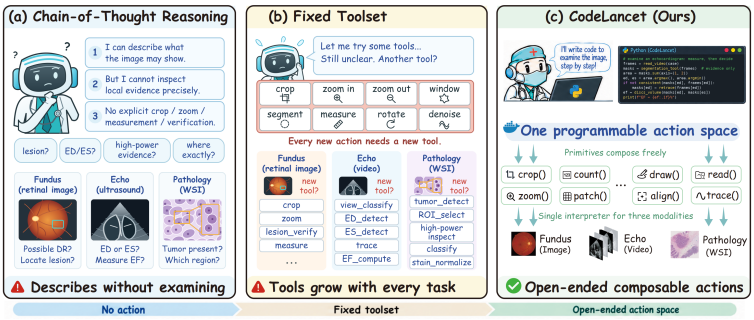

Figure 1: **Three ways for a VLM to act on a medical image.** (a) Chain-of-thought describes an examination without conducting one. (b) A fixed toolset acts, but each task needs new tools whose workflow the model follows. (c) CodeLancet executes code on the original pixels: operations compose freely across modalities, and each act is an inspectable decision.

Nor would a threshold correct this: the numeric estimates of all six frontier models are essentially uncorrelated with the reference (Pearson $r \leq 0.09$; Figures 3 and 4). Two structural causes underlie this. A single forward pass offers no operation for zooming, cropping, registration or measurement, so no quantity is explicitly computed; and clinical reports record what a physician found but not how it was found, so the procedure is largely absent from training data. A model that cannot examine falls back on the population prior and misses the patients whose diagnosis needs examining.

Two lines of work try to restore the procedure (Figure 1). The first prompts or trains the model to reason at length before answering (Wei et al., 2022) (Figure 1a), but such reasoning describes an examination without conducting one. The rationale need not reflect the computation that produced the answer (Turpin et al., 2023): a landmark may be described that was never located, and medical VLMs often state findings the image does not support (Xia et al., 2024). Because a rationale records no action on the image and no observation in return, it offers no interaction to learn from. The second line lets the model act through a *fixed toolset* (Yao et al., 2023; Zheng et al., 2026; Li et al., 2024) (Figure 1b), which fixes in advance which operations exist; training on it teaches the model to follow the procedure its tools encode rather than internalize it. We confirm both limits: given tools, frontier models improve only where an instrument supplies the evidence, and given the physician's written procedure, they comply without judgment (Section 4.1); a fixed-toolset agent trained identically to ours falls near chance outside its toolset (Section 4.2). Tools alone do not confer examination, nor does text prescribe it; it must be learned, in an action space that does not fix it in advance.

Our key insight is that the physician's procedure is a sequence of *acts* on the image, and that what a VLM lacks is a medium in which such acts can be composed, executed and learned. Executable code provides one (Figure 1c), because it does not fix the granularity of an act. The policy can reach below any single step to inspect and recompute an intermediate result, reach beyond the anticipated steps to compose an operation no designer foresaw, and lengthen or shorten the examination as the case requires, treating perception itself as an act that chooses where and how finely to look in space, time and magnification, all within one programmable space that spans images, videos and gigapixel slides. A fixed toolset limits all three: its steps are opaque to the policy, its coverage is fixed at design time, and its effort does not adapt to the case. Because every act is an executed statement, an examination written in code also decomposes into discrete decisions, each of which can be checked against the physician's procedure and credited on its own; the procedure becomes learnable rather than merely executable. As the lancet is the instrument through which a physician's hand acts on a patient, code is the instrument through which a model acts on an image. The question is therefore no longer whether a VLM can recognize a diagnosis but **whether it can learn to examine the way a physician does**.

We present CodeLancet, a 4B agent that examines medical images through executable code, trained in two stages (Figure 2). Since few datasets record how a diagnosis was reached, physicians first write their reading as structured *skills*; a teacher turns each skill into code-executing demonstrations, and a rubric keeps only those that follow every prescribed step. Stage 1 fine-tunes the agent on them and reinforces it with GRPO on the outcomes of its own examinations. Yet outcome reward makes the trajectory the unit of credit: wrong steps in a correct examination are reinforced along with the rest,

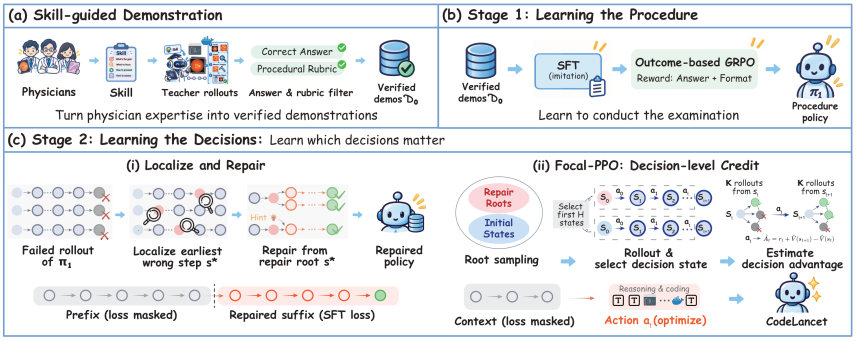

Figure 2: **How CodeLancet learns to examine.** (a) A teacher demonstrates physician-written skills; only trajectories whose answer and prescribed steps are verified are kept. (b) Stage 1 learns the procedure. (c) Stage 2 learns the decisions through localize-and-repair and Focal-PPO.

and a group whose rollouts all succeed carries no signal. The result is step-level hollow correctness: a blinded judge finds critical wrong steps in more than a third of the agent's *correct* examinations. Stage 2 therefore makes the decision the unit of learning: the agent is fine-tuned to repair the earliest wrong step a judge localizes in each failed examination, and *Focal-PPO* credits each decision by the change in expected outcome it causes, estimated from rollouts branched around the decisions after each localized failure, which stays informative even when every rollout of a case ends alike.

We evaluate on six public benchmarks spanning three modalities and three kinds of clinical task against six frontier VLMs. Our contributions are threefold. (1) *We identify hollow correctness at two levels and trace each to its cause*: at the level of the answer, to the absence of an action space in which to obtain evidence; at the level of the step, to trajectory-level credit that cannot reach the decision that earned it. (2) *We present CodeLancet*, which internalizes physician-authored skills coarse to fine: outcome-level optimization of the examination, then decision-level credit assignment through localize-and-repair and Focal-PPO, which cut wrong steps in correct examinations from 36% to 10% where a further stage of outcome reward is nearly flat (Sections 2.3 and 4.3). (3) *We show that a 4B agent examining through code outperforms six frontier VLMs on every headline metric*, with a margin that widens as the answer depends more on examination than on recognition; at comparable accuracy it detects 72.5% of failing hearts against at most 26.3%. The gain comes from both training and the code action space: trained identically, a fixed toolset approaches code where it anticipates the operation, falls near chance where it does not, and catches far fewer of the faulty tool outputs that code detects and repairs. Looking closely only where the diagnosis is decided, the agent reaches full-resolution evidence at about 1% of the visual cost of a detailed whole-slide read and 15% of an echocardiogram's frames, and blinded specialists endorse 80–86% of its steps (Sections 3.2–4.4).

2 METHOD

Problem formulation. Given a case x (a fundus photograph, echocardiogram or whole-slide image, with metadata) and a clinical question q, the agent conducts an examination in a sandboxed Python interpreter. Media enter through two parameterized readers whose arguments the policy sets at every call, the time window, frame rate and resolution of a video and the magnification and grid cells of a slide; they return raw pixels and prescribe no procedure, so the policy can survey coarsely before examining a window or region at full resolution. For echocardiography, a left-ventricular segmentation tool also proposes masks but no ejection fraction. These *instruments* supply evidence, never an answer: every quantity the answer depends on is computed in code (Appendix C). A *decision* $a_t \sim \pi_\theta(\cdot \mid s_t)$ is one assistant turn that reasons and then invokes the interpreter or an instrument, or answers; o_t is its output, including images, and $s_t = (x, q, a_1, o_1, \ldots, a_{t-1}, o_{t-1})$. An examination is a trajectory $\tau = (a_1, o_1, \ldots, a_T, o_T)$ ending in an answer $\hat{y}(\tau)$ with terminal reward $r(\tau)$. The policy is initialized from Qwen3.5-4B (Qwen Team, 2026).

2.1 SKILL-GUIDED DEMONSTRATION

For each task (Figure 2a), board-certified subspecialists author an *examination skill* $\mathcal{S}$ that specifies the clinical goal, anatomical anchors, diagnostic criteria, an investigative workflow and an evidential discipline: every finding is confirmed in the original pixels, and every instrument output is checked rather than trusted (Appendix C). A teacher π_{T} receives $\mathcal{S}$ and examines each training case in the student's sandbox. Since the answer alone would admit trajectories that reach the right grade while skipping the step that justifies it, a judge J also scores a rubric $c_{\mathcal{S}}(\tau) \in \{0,1\}$ from the executed trajectory and its images, without the label, equal to 1 only if every applicable evidence-bearing step of $\mathcal{S}$ was executed with visible evidence. With $\mathrm{ok}(\tau) = \mathbf{1}[\hat{y}(\tau) = y^\star]$ for categorical items and $\mathbf{1}[|\hat{e}(\tau) - e^\star| \le 5]$ for the numeric ejection fraction (in points), the demonstration set is $\mathcal{D}_0 = \{\tau \sim \pi_{\mathrm{T}}(\cdot \mid x, q, \mathcal{S}) : \mathrm{ok}(\tau) = 1 \wedge c_{\mathcal{S}}(\tau) = 1\}$. The rubric removes 31% of the answer-correct teacher rollouts, most often for asserting a check that was never run (Appendix D).

2.2 STAGE 1: LEARNING THE PROCEDURE

The student is fine-tuned on $\mathcal{D}_0$ with $\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{\tau\sim\mathcal{D}_0} \sum_t \log \pi_\theta(a_t \mid s_t)$ and then optimized with GRPO (Shao et al., 2024) on the outcome of its own examinations (Figure 2b), with reward $r(\tau) = 0.9\, r_{\mathrm{task}}(\tau) + 0.1\, r_{\mathrm{fmt}}(\tau)$, where $r_{\mathrm{fmt}} \in \{0,1\}$ rewards a well-formed answer and $r_{\mathrm{task}}(\tau)$ is $\mathbf{1}[\hat{y}(\tau) = y^\star]$ for categorical items and $1 - \min(|\hat{e}(\tau) - e^\star|, \delta)/\delta$ for the numeric EF, which decays linearly to zero at $\delta = 10$ points. Advantages are normalized within each group of G rollouts, $\hat{A}_i = (r_i - \bar{r})/\sigma_r$, and groups with $\sigma_r = 0$ are dropped. Two choices keep the reward from paying for priors: option letters are redrawn for every rollout and batches are balanced over categories, and the EF reward is placed on the estimated value rather than on its category. The result is π_1.

2.3 STAGE 2: LEARNING THE DECISIONS

In Stage 1 every decision of rollout τ_i shares the advantage $\hat{A}_i$, and a group whose rewards coincide contributes none; wrong decisions in a rollout with positive advantage are reinforced along with the rest, and π_1 plateaus with critical wrong steps in 36% of its *correct* test examinations (Section 4.3). Stage 2 makes the decision, rather than the trajectory, the unit of learning (Figure 2c).

(i) Localize and repair. For each training case that π_1 fails, J receives $\mathcal{S}$, the trajectory and its images, but not the label, and returns the earliest decision $t^\star$ that the visible evidence shows to be wrong, or abstains; $s^\star = s_{t^\star}$ is a *repair root*. A continuation $\tau' \sim \pi_1(\cdot \mid s^\star)$ is accepted when the completed examination meets both conditions of $\mathcal{D}_0$. If none is, J supplies a *hint* naming the next thing to verify but never what is present; the hint conditions sampling only and is deleted before training, so an accepted action must pass the audit from the prefix alone. Training on the new decisions ($t \ge t^\star$) of accepted continuations, mixed with audited replays of π_1's correct examinations, yields π_2.

(ii) Focal-PPO: decision-level credit. Roots are the repaired roots together with initial states of training cases. From each root, π_2 rolls out one examination, and for each of its first H decisions the states before and after the decision are valued by K independent continuations to termination:

$$\hat{V}(s) = \frac{1}{K}\sum_{k=1}^{K} r(\tau_k), \quad \tau_k \sim \pi_2(\cdot \mid s), \qquad \hat{A}(s_t, a_t) = r_t + \hat{V}(s_{t+1}) - \hat{V}(s_t), \tag{1}$$

where r_t is the terminal reward if a_t ends the episode and 0 otherwise. We maximize the clipped surrogate (Schulman et al., 2017) token-wise within each decision,

$$\mathcal{J}(\theta) = \mathbb{E}_{(s_t,a_t)}\, \frac{1}{|a_t|}\sum_{j=1}^{|a_t|} \min\Big(\rho_{t,j}\,\hat{A}(s_t,a_t),\ \mathrm{clip}\big(\rho_{t,j}, 1-\epsilon_{\mathrm{lo}}, 1+\epsilon_{\mathrm{hi}}\big)\hat{A}(s_t,a_t)\Big), \tag{2}$$

where $\rho_{t,j}$ is the probability ratio of the decision's j-th token under π_θ and the frozen π_2; there is no learned critic, group normalization or KL term. $\hat{A}$ is an unbiased Monte-Carlo estimate of the decision's advantage under π_2 (Kazemnejad et al., 2024), which can be non-zero even when every rollout of a case ends alike (Appendix D). What makes it focal is where it is spent: at the decisions that follow a localized failure; the same estimator from initial states alone loses most of the accuracy gain (Table 3). The result is CodeLancet (H=3, K=4; Appendix D).

Table 1: **Main results.** CodeLancet (4B) versus six frontier VLMs answering directly from the same media. Best in **bold**; ↓ lower is better; Δ: CodeLancet minus the best frontier value of each metric.

	DR grading			Ejection fraction					Melanoma relapse			
Model	Acc	QWK	Ref. err.↓	Acc	Bal. acc	Macro F1	MAE↓	R^2	F1	PPV	Spec	False alert.↓
GPT-4o	59.6	45.1	27.2	22.6	34.0	20.9	10.47	−0.33	50.9	40.0	80.0	16.8
GPT-5.4 Mini	61.9	41.9	30.8	48.4	34.3	32.1	13.59	−0.92	47.7	35.9	75.7	20.4
Gemini 2.5 Flash	55.5	20.9	36.9	60.5	34.3	34.1	12.98	−1.10	44.2	30.2	63.5	30.6
Claude Sonnet 4.6	63.3	46.1	25.5	68.0	32.3	30.7	12.72	−0.47	45.7	31.3	64.4	29.9
GPT-5.6 Luna	71.3	55.3	21.6	66.2	33.0	32.9	10.66	−0.24	34.2	23.4	60.2	33.4
Gemini 3.5 Flash Lite	75.2	57.9	17.8	69.5	37.2	35.0	12.14	−0.91	35.7	24.1	58.2	35.1
CodeLancet (ours, 4B)	**81.4**	**74.9**	**9.5**	**69.7**	**62.1**	**54.6**	**7.32**	**0.38**	**53.2**	**45.0**	**84.8**	**12.8**
Δ vs. best frontier	+6.2	+17.0	−8.3	+0.2	+24.9	+19.6	−3.15	+0.62	+2.3	+5.0	+4.8	−4.0

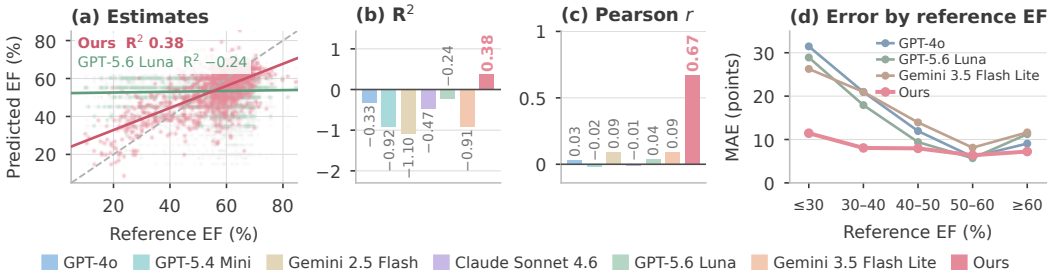

Figure 3: **Numeric ejection fraction** ($n = 1{,}277$). (a) CodeLancet and GPT-5.6 Luna, the frontier model with the highest R^2; (b,c) agreement with the reference; (d) error by reference-EF band.

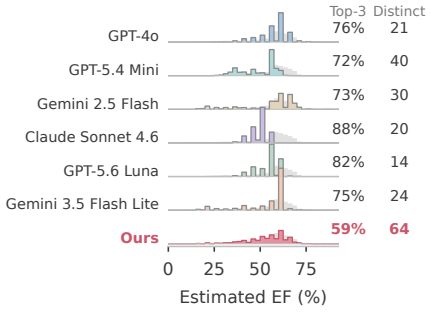

Figure 4: **Distribution of numeric EF estimates** over the reference (grey).

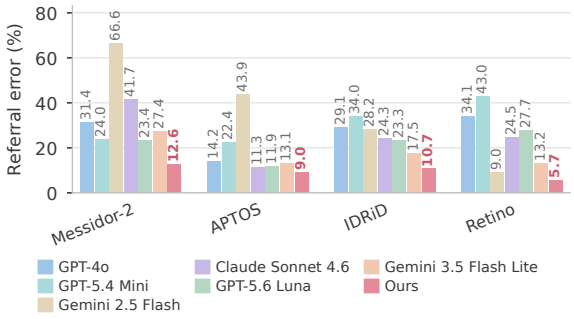

Figure 5: **DR referral-decision error by held-out cohort.**

3 Experiments

3.1 Setup

Tasks and metrics. Six public benchmarks span three modalities and three kinds of clinical task: *ordinal grading* of diabetic retinopathy (DR) on 2,296 held-out fundus photographs from four cohorts (Decencière et al., 2014; Asia Pacific Tele-Ophthalmology Society, 2019; Porwal et al., 2018; Wang et al., 2023); *quantitative measurement* of ejection fraction (EF), as a three-category item and as a numeric estimate, on the 1,277 EchoNet-Dynamic test loops (Ouyang et al., 2020); and *prognosis* of five-year melanoma relapse on 1,322 VisioMel whole-slide images (DrivenData, 2023). Metrics are those of Table 1; datasets are described in Appendix B and full metric panels given in Appendix F.

Baselines. Six frontier VLMs (Table 1; Hurst et al., 2024; OpenAI, 2026a; Comanici et al., 2025; Anthropic, 2026; OpenAI, 2026b; Google DeepMind, 2026) answer directly from the same media, question and answer format (prompts in Appendix E).

Table 2: **Attribution ladder.** Frontier models answering directly, with CodeLancet's sandbox and instruments (+Tools), and additionally with the skill in the system prompt (+Tools +Skill).

			DR grading		Ejection fraction			Relapse	
Model	Condition	Skill	Acc	Ref. err↓	Bal. acc	MAE↓	R^2	PPV	F1
GPT-5.6 Luna	Direct	–	71.3	21.6	33.0	10.66	−0.24	23.4	34.2
	+Tools	–	57.3	29.6	42.2	18.58	−1.88	42.4	50.8
	+Tools +Skill	✓	60.7	29.0	47.1	14.91	−1.07	42.3	51.7
Gemini 3.5 Flash Lite	Direct	–	75.2	17.8	37.2	12.14	−0.91	24.1	35.7
	+Tools	–	72.7	21.4	34.5	14.52	−1.16	44.4	51.5
	+Tools +Skill	✓	73.8	20.8	44.7	12.73	−0.69	44.1	51.9
Claude Sonnet 4.6	Direct	–	63.3	25.5	32.3	12.72	−0.47	31.3	45.7
	+Tools	–	57.4	29.8	52.2	9.52	0.04	29.2	44.2
	+Tools +Skill	✓	63.7	28.5	56.8	10.77	−0.27	34.7	48.0
CodeLancet (4B)	Trained	–	**81.4**	**9.5**	**62.1**	**7.32**	**0.38**	**45.0**	**53.2**

3.2 Main results

Numeric estimates track the ventricle, not the population. The most accurate frontier models reach 66–70% raw EF accuracy by predicting the majority class and detect only 11–26% of hearts with reduced EF; at comparable accuracy (69.7%), CodeLancet detects 72.5% (Table 8). Frontier numeric estimates are uncorrelated with the reference (Pearson $r \leq 0.09$; R^2 from -1.10 to -0.24), and 72–88% of them fall within each model's three most frequent 5-point bins; CodeLancet, which computes the ventricular volume in code, attains an MAE of 7.3 points and $R^2 = 0.38$ with 64 distinct values against 14–40 (Figures 3a–c and 4). The margin is largest where the population prior is least informative: at reference EF of 30% or less, frontier error is two to three times each model's overall MAE, whereas CodeLancet stays within 6–12 points in every band (Figure 3d).

Lesion-level verification nearly halves referral errors. CodeLancet resolves proliferative signs first in 91% of examinations and confirms candidate lesions at native resolution; its referral-decision error of 9.5% against 17.8% for the best frontier model is the lowest on every cohort (Figures 5 and 10).

Multiscale traversal sharpens relapse alerts. Surveying the slide at low power and verifying candidate regions at high power, CodeLancet raises the PPV of a relapse alert to 45.0% (frontier 23.4–40.0%) and lowers the false-alert rate to 12.8% (16.8–35.1%; Table 1).

4 Analysis and Ablations

4.1 Where does the gain come from: tools, procedure, or training?

The attribution ladder (Table 2) gives GPT-5.6, Gemini 3.5 and Claude the agent's sandbox and instruments (+Tools) and also the skill (+Tools +Skill), to test whether tools, the written procedure or training makes a model examine; CodeLancet receives no skill text at inference.

Tools help where an instrument supplies the evidence; on retinopathy they make frontier models worse. On retinopathy all three models fell below their direct-answer accuracy once given tools, GPT-5.6's referral error rising from 21.6% to 29.6% (Table 2), and GPT-5.6 and Gemini called the interpreter only 2–3 times per case. On echocardiography the models rarely questioned the segmentation output, treating it as an oracle rather than as evidence to be verified (Tables 10 and 6).

The written procedure yields compliance, not judgment. The skill text raised balanced EF accuracy for all three models to 44.7–56.8%, yet none recovered its direct-answer referral error on retinopathy. Claude followed the prescribed echocardiography sequence in 89.8–94.8% of cases, and still fell short; the other two followed it in as few as 12.5% of cases (Table 10).

Tool access and procedural prompting do not close the gap. Given the same sandbox, instruments and skill, no frontier configuration reaches CodeLancet on any metric; the best trails by 6.2 points of DR accuracy, 5.3 of balanced EF accuracy and 1.3 of relapse F1 (Table 2), and the gap is widest where the procedure demands verification and measurement.

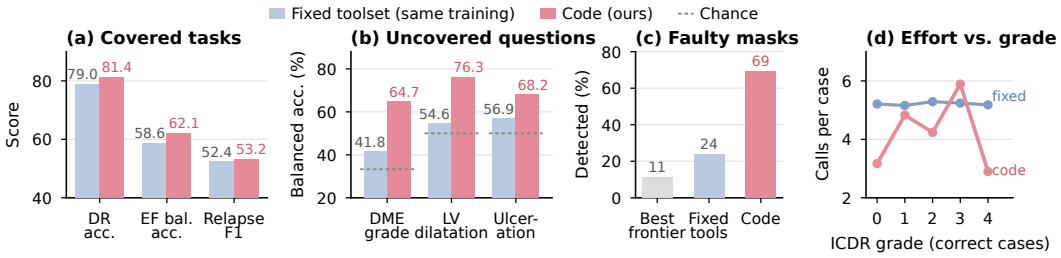

Figure 6: **Code versus a fixed toolset under identical training.** (a) Tasks the toolset covers; (b) held-out questions outside it; (c) faulty masks detected; (d) calls per correct retinal examination.

Table 3: **Two stages of learning.** Performance after each stage and for Stage 2 alternatives. Defect: share of test examinations with a critical defect; Budget: sampling budget relative to Focal-PPO.

	DR grading		Ejection fraction		Relapse	Defect↓	Budget
Checkpoint	Acc	Ref. err↓	Bal. acc	MAE↓	F1	(%)	(rel.)
Stage 1, SFT on $\mathcal{D}_0$	73.6	14.3	49.7	9.63	46.3	52.4	–
Stage 1, + outcome GRPO (π_1)	78.4	12.9	57.6	8.37	50.7	42.6	–
Stage 2, + localize-and-repair SFT (π_2)	79.7	10.8	60.2	7.81	52.1	27.3	–
Stage 2, + Focal-PPO (CodeLancet)	**81.4**	**9.5**	**62.1**	**7.32**	**53.2**	**13.1**	1.0×
Alternatives							
Second GRPO stage from π_2, instead of Focal-PPO	79.9	11.2	60.6	7.86	51.9	26.1	2.6×
Focal-PPO from initial states only (no repair roots)	80.3	10.6	60.7	7.61	52.5	19.8	1.0×
Repair SFT without hints (replaces π_2)	79.3	11.5	59.5	7.98	51.6	30.4	–

4.2 Why code rather than a fixed toolset?

To separate the action space from training, we train a fixed-toolset variant of CodeLancet with the same backbone, skills, teacher, rubric, training stages and rollout budget, replacing the interpreter by one dedicated tool for each operation the skills prescribe (Appendix C), following the design of prior medical agents (Appendix A), and compare the two agents in Figure 6.

A fixed tool's output can be rejected but not repaired. The fixed-toolset agent trails CodeLancet even on the tasks its toolset covers: 79.0% against 81.4% DR accuracy, 58.6% against 62.1% balanced EF accuracy, and 52.4 against 53.2 relapse F1 (Figure 6a). The gap is largest on ejection fraction, whose tool returns a value the agent can accept or reject but not recompute. Given a faulty mask, the fixed-toolset agent detects the fault in 24% of cases against 69% for CodeLancet; unable to redraw the boundary, it reaches an MAE of 15.6 on them, close to the 16.4 of adopting the mask, whereas CodeLancet, which corrects over half of the detected faults in code, reaches 9.8 (Figure 6c; Table 11).

Outside its toolset, the fixed-toolset agent can recognize but not examine. Neither agent is trained on three held-out questions (macular edema grade, left-ventricular dilatation and ulceration; Appendix B), and each receives only a one-sentence criterion. The fixed-toolset agent attains 41.8%, 54.6% and 56.9% balanced accuracy, within nine points of chance, whereas CodeLancet composes the missing operation in code and attains 64.7%, 76.3% and 68.2% (Figure 6b). Answering these questions with a toolset requires new tools; with code, the interface is unchanged.

A fixed toolset spends nearly the same effort on every case. The fixed-toolset agent issues a near-constant five calls per retinal examination at every grade, mostly in the order its tools are listed. CodeLancet's effort instead follows the decision each case requires: correct examinations close in 2.9 calls on proliferative disease once neovascularization is confirmed, but take 5.9 on severe non-proliferative disease, where the 4-2-1 rule must be counted (Figure 6d). Training instils the procedure in either action space; in code, the learned procedure can also correct its own evidence, reach operations the toolset lacks and scale its effort to the case.

4.3 Learning the procedure, then the decisions

We ask what each stage contributes, whether Stage 2 corrects the step or only the answer, and whether its gain comes from decision-level credit or from more rollouts (Table 3, Figure 7).

Each stage improves every metric, and Stage 2 drives most of the drop in defect rate. Outcome GRPO lifts DR accuracy from 73.6% to 78.4% and EF balanced accuracy from 49.7% to 57.6%,

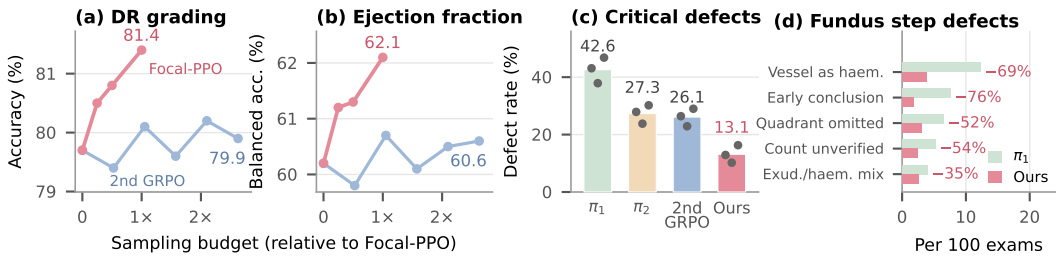

Figure 7: **Stage 2 from π_2.** (a,b) A second GRPO stage versus Focal-PPO as the budget grows; (c) critical-defect rate by checkpoint (dots: tasks); (d) fundus step defects before and after Stage 2.

Table 4: **Where correction is learned.** Code self-repair (failed blocks re-run successfully within two turns; cases correct despite a code error) and adjudication of faulty masks (Table 6) by checkpoint.

	Code self-repair		Instrument adjudication		
Checkpoint	Repair rate (%)	Correct despite error (%)	Detected (%)	Corrected (%)	EF bal. acc / DR acc
Base model + tools (no training)	30.8	19.8	5.7	2.9	35.8 / 58.2
Stage 1, SFT on $\mathcal{D}_0$	53.6	38.9	38.2	29.5	49.7 / 73.6
Stage 1, + outcome GRPO (π_1)	66.8	55.7	52.4	44.9	57.6 / 78.4
Stage 2, + localize-and-repair SFT (π_2)	74.9	64.1	60.3	51.2	60.2 / 79.7
Stage 2, + Focal-PPO (CodeLancet)	**84.7**	**77.3**	**68.9**	**61.8**	**62.1 / 81.4**
Without demonstrations					
Outcome GRPO from the base model	39.4	26.4	9.4	5.1	40.9 / 65.2

then plateaus (Table 3); applied to the base model without demonstrations, it reaches only 65.2% and 40.9% (Table 4), so the demonstrations supply the procedure that outcome reward refines. Localize-and-repair SFT adds 1.3 and 2.6 points, and Focal-PPO a further 1.7 and 1.9; three-quarters of the total reduction in critical defects (52.4% to 13.1%) comes from Stage 2.

Correction lands on the step, not only on the answer. A blinded judge auditing every test examination against the skill finds a critical defect in 42.6% of π_1's examinations and 13.1% of CodeLancet's; among *correct* examinations the rate falls from 36% to 10%. Stage 2 thus removes most of the hollow correctness at the level of the step, and on fundus photographs the defect types most often localized in Stage 2 become 35–76% rarer per examination (Figure 7c,d).

A second stage of outcome reward is nearly flat; decision-level credit is not. From π_2, a second GRPO stage adds only 0.2 and 0.4 points of DR and balanced EF accuracy with 2.6 times the sampling budget of Focal-PPO, which reaches 81.4% and 62.1% and still improves (Figure 7a,b). What differs is the signal: most GRPO groups on categorical items keep no reward variance and hence no gradient, whereas decision-level estimates can remain informative on such cases (Appendix D).

The gain lies in where credit is spent. With estimator and budget fixed, Focal-PPO from initial states loses most of its accuracy gain (80.3% and 60.7% against 81.4% and 62.1%), as do depth-matched random roots in the same failed examinations (Figure 9). Hints make the wrong step repairable: without them, 39% of roots stay unrepaired and about a third of the repair gain is lost (Table 3).

4.4 Inside the learned examination

Code enables active perception at a fraction of the visual cost. On whole-slide images, CodeLancet surveys the slide at low power and reserves high magnification for the few regions that decide the diagnosis, at about 1% of the visual cost of reading the whole slide at high power; frontier models answering directly see only the overview (Figure 8). It typically reads an echocardiogram the same way: a sparse pass locates a cardiac cycle and a dense re-read locks end-diastole and end-systole, using about 15% of the native frames, concentrated on the phases that decide the measurement.

Fundus grading needs no learned instrument. With the interpreter as its only tool, the agent writes the perception itself (Appendix H), yet its referral error is 9.5% against 20.8% for the best frontier configuration with the identical tool and 17.8% for the best direct answer (Table 2).

The examination is worth more than the instrument, and the agent verifies it. Without the segmentation tool at inference, the agent traces the boundary in code and retains 58.1% balanced accuracy (from 62.1%), ahead of every frontier configuration (Table 5). Given a faulty mask (Dice

Table 5: **Removing the segmentation tool** at inference (*zero-shot*) or throughout (*retrained*).

	Ejection fraction			Relapse (control)		
System	Bal. acc	MAE↓	R^2	F1	PPV	False alert↓
Best direct frontier value, per metric	37.2	10.47	−0.24	50.9	40.0	16.8
Best +Tools +Skill frontier value, per metric	56.8	10.77	−0.27	51.9	44.1	12.9
CodeLancet, code-only (zero-shot)	58.1	8.72	0.23	**53.2**	**45.0**	**12.8**
CodeLancet, code-only (retrained)	60.4	7.66	0.33	52.9	44.5	13.0
CodeLancet, full instruments	**62.1**	**7.32**	**0.38**	**53.2**	**45.0**	**12.8**

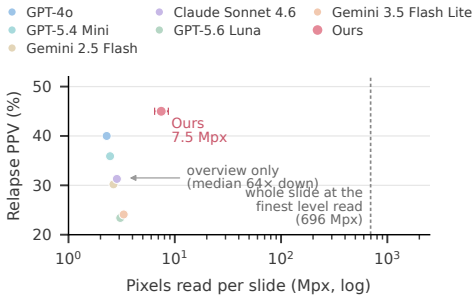

Figure 8: **Visual budget on whole-slide images** (CodeLancet: median and IQR).

Table 6: **Adjudication of faulty instrument output on echocardiography.** A proposal is faulty when its mask has Dice < 0.7 to the expert trace, and detected when the agent rejected, re-ran or overrode it; MAE on the faulty cases. Frontier models: +Tools +Skill.

			MAE↓	
System	Faulty (%)	Detected (%)	Adopted	Answer
Claude Sonnet 4.6	15.9	11.3	17.6	16.1
GPT-5.6 Luna	13.7	5.1	16.2	18.7
Gemini 3.5 Flash Lite	14.1	7.8	17.5	17.0
CodeLancet	14.3	**68.9**	16.4	**9.8**

Table 7: **Blinded physician evaluation.** Endorsement (%; score ≥4 of 5) of 150 trajectories by three subspecialists per modality, mean pairwise quadratic-weighted κ_w and 95% bootstrap intervals.

		Case level		Step level	
Modality		Evidence grounding	Workflow concordance	Step necessity	Step correctness
Fundus	Endorsement (%)	82.7 [74.0, 90.0]	80.0 [70.0, 88.0]	83.4 [77.1, 88.7]	84.6 [80.1, 88.8]
	Agreement (κ_w)	0.78 [0.60, 0.89]	0.74 [0.49, 0.86]	0.83 [0.76, 0.87]	0.71 [0.62, 0.78]
Echo	Endorsement (%)	86.0 [78.0, 92.7]	84.0 [75.3, 91.3]	85.9 [81.7, 89.9]	82.7 [78.6, 86.8]
	Agreement (κ_w)	0.76 [0.56, 0.85]	0.80 [0.69, 0.86]	0.73 [0.67, 0.77]	0.71 [0.64, 0.76]
Pathology	Endorsement (%)	84.7 [75.3, 92.0]	81.3 [71.3, 89.3]	81.2 [76.6, 85.2]	79.8 [74.8, 84.6]
	Agreement (κ_w)	0.75 [0.64, 0.84]	0.70 [0.54, 0.82]	0.77 [0.71, 0.82]	0.74 [0.66, 0.80]

< 0.7; 14.3% of cases), CodeLancet detects 69% of these faults by checking the area curve against frames it has read, whereas frontier models given the same tool and told to verify detect 5–11% (Table 6); detection is learned, rising from 6% in the base model, mostly in Stage 1 (Table 4).

Physicians endorse most steps of the learned examination. Three board-certified subspecialists per modality reviewed 150 trajectories blinded to label and correctness (Appendix I); endorsement was 80–86% at both the case and the step level (mean pairwise κ_w 0.70–0.83; Table 7).

5 Discussion, Limitations and Conclusion

We presented CodeLancet, a 4B agent that examines medical images through executable code and internalizes physician-authored skills from outcome-level reinforcement to decision-level credit with Focal-PPO. It outperforms all six frontier VLMs we evaluate on every headline metric, most where the answer depends on examination rather than recognition. Two controlled experiments locate the gain: a fixed toolset trained identically is held back by opaque steps, fixed coverage and non-adaptive effort, and a further stage of outcome reward is nearly flat where decision-level credit still improves. Correct answers also rest on sounder examinations, as critical wrong steps fall from 36% to 10% of them, replacing hollow correctness with conclusions whose derivation can be re-executed. Limitations remain: evaluation is retrospective, pending prospective and reader studies, and each new task requires physician-authored skills, whose cost bounds scaling. The examination policy, rather than any single instrument, is the durable asset: instruments can be removed without retraining the judgment that checks them, and each new skill extends the same action space to a new task.

AI USE STATEMENT

In accordance with the ICLR policy on the use of AI, we disclose the following. We used generative AI tools to generate synthetic training data: a teacher model (Qwen3.8-Flash-Next) produces the demonstration trajectories, and Gemini 3.6 Flash serves as the judge that audits them, localizes the first erroneous decision in failed examinations, verifies the repaired continuations used for localize-and-repair fine-tuning, and scores the defect rates of test examinations (Section 2; Appendix D). We also used large language model assistants to polish the wording and grammar of text written by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this submission.

ETHICS STATEMENT

This study analysed exclusively publicly available, de-identified datasets (Messidor-2, IDRiD, APTOS 2019, MedFMC/Retino, EchoNet-Dynamic and VisioMel) under their original access terms; no new patient data were collected. The physician evaluation involved expert review of model outputs and public images only. CodeLancet is a research system evaluated retrospectively and is not intended for clinical use without prospective validation.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix C specify the action space, sandbox semantics and instrument interfaces; Section 2 gives all training stages and Appendix D their hyperparameters and protocols; Section 3.1 and Appendix B give the evaluation sets, label mappings and per-cohort sample sizes; Appendix E gives the baseline prompts. Code and model weights will be released.

REFERENCES

Anthropic. Claude Sonnet 4.6 System Card. https://www.anthropic.com/claude-sonnet-4-6-system-card, February 2026. Accessed 2026-09-04.

Asia Pacific Tele-Ophthalmology Society. APTOS 2019 blindness detection. Kaggle competition, https://www.kaggle.com/c/aptos2019-blindness-detection, 2019.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pp. 231–234, 2014.

DrivenData. VisioMel Challenge: Predicting melanoma relapse. https://www.drivendata.org/competitions/148/visiomel-melanoma/, 2023. Accessed 2023-04-01.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Jeffrey E Gershenwald, Richard A Scolyer, Kenneth R Hess, Vernon K Sondak, Georgina V Long, Merrick I Ross, Alexander J Lazar, Mark B Faries, John M Kirkwood, Grant A McArthur, et al. Melanoma staging: evidence-based changes in the american joint committee on cancer eighth edition cancer staging manual. *CA: a cancer journal for clinicians*, 67(6):472–492, 2017.

Google DeepMind. Gemini 3.5 Flash-Lite model card. https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/, 2026. Accessed 2026-09-24.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. *arXiv preprint arXiv:2410.01679*, 2024.

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, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging. *European Heart Journal-Cardiovascular Imaging*, 16(3):233–271, 2015.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8745–8760, 2024.

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.

OpenAI. Introducing GPT-5.4 mini and nano. `https://openai.com/index/introducing-gpt-5-4-mini-and-nano/`, 2026a. Accessed 2026-09-24.

OpenAI. GPT-5.6 system card. `https://deploymentsafety.openai.com/gpt-5-6`, 2026b. Accessed 2026-09-24.

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, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026. Alibaba Group. Accessed 2026-09-04. Official model card for Qwen3.5-4B available at `https://huggingface.co/Qwen/Qwen3.5-4B`.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Scientific Data*, 10(1):574, 2023.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. *arXiv preprint arXiv:2402.01030*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Charles P Wilkinson, Frederick L Ferris III, Ronald E Klein, Paul P Lee, Carl David Agardh, Matthew Davis, Diana Dills, Anselm Kampik, Rangasamy Pararajasegaram, Juan T Verdaguer, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37:140334–140365, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

A Related work

Medical VLMs. Medical image AI has moved from task-specific classifiers (Gulshan et al., 2016; Esteva et al., 2017) to foundation models that learn transferable structure from unlabelled retinal images (Zhou et al., 2023), and generalist systems that unify images, clinical language and biomedical knowledge have been proposed (Moor et al., 2023). Scale has raised recognition accuracy, but the model still reads the image once and answers, and chain-of-thought rationales can be unfaithful to the computation that produced the answer (Turpin et al., 2023). We treat this as an action-space problem rather than an explanation problem.

Medical agents and tool use. A second line places the model in an agent loop that calls tools (Yao et al., 2023), applied to image inspection and medical image analysis (Zheng et al., 2026; Li et al., 2024). These systems act through a designer-fixed set of tools, whereas executable code composes primitives freely (Wang et al., 2024). Reinforcement learning for language-model reasoning assigns credit from outcome rewards over whole trajectories or from step-level value estimates along on-policy trajectories (Shao et al., 2024; Yu et al., 2026; Kazemnejad et al., 2024). CodeLancet differs from both lines in that the examination is composed by the policy rather than by a designer, and decision-level credit is estimated where a label-blind judge localizes the first error of a failed examination, with states restored by re-executing code in the sandbox.

B Datasets

Diabetic retinopathy. Four public colour fundus photography datasets were used for held-out evaluation ($n = 2{,}296$: Messidor-2, 350; APTOS 2019, 733; IDRiD, 103; Retino, 1,110), spanning populations in India, France and China and heterogeneous camera systems. Messidor-2 (Decencière et al., 2014) contains macula-centred photographs from French ophthalmology departments, used with five-level ICDR grades (Wilkinson et al., 2003). IDRiD (Porwal et al., 2018) provides high-resolution images with pixel-level lesion annotations and image-level grades from Nanded, India. APTOS 2019 (Asia Pacific Tele-Ophthalmology Society, 2019) comprises images from the Aravind Eye Hospital

system graded by trained graders. Retino is the fundus subset of MedFMC (Wang et al., 2023), 1,392 photographs from Shanghai Tenth People's Hospital (one per patient; mean resolution 2,736×1,824) with five-level DR labels, of which 1,110 form the held-out evaluation set. Five-level labels were collapsed to three classes (ICDR 0 → R0; 1 → R1; 2–4 → R2, referable). Training cases were drawn from the training partitions of these sources, disjoint at the case level from the evaluated images.

Echocardiography. EchoNet-Dynamic (Ouyang et al., 2020) contains 10,030 apical-four-chamber videos from Stanford University Medical Center with expert-traced end-diastolic and end-systolic segmentations and a method-of-discs reference EF. The official test partition ($n = 1{,}277$) was used, with categories reduced ($\leq$40%; 12.5% of loops), mildly reduced (40–50%; 9.8%) and preserved ($\geq$50%; 77.7%). The agent computed a single-plane method-of-discs EF from the apical-four-chamber view.

Melanoma relapse. VisioMel (DrivenData, 2023) provides whole-slide images of primary cutaneous melanoma excisions from a French multicentre cohort, each paired with the baseline clinical metadata released with the dataset, with five-year relapse as the binary endpoint ($n = 1{,}322$ evaluated cases). Predicting relapse from the primary excision is clinically important because adjuvant therapy depends on recurrence risk (Gershenwald et al., 2017).

Tasks and answer formats. Each system receives identical items. *Diabetic retinopathy* is posed as a five-option multiple-choice question over the ICDR grades (no DR; mild, moderate and severe non-proliferative DR; proliferative DR) with the option letters shuffled per item; the reference is the dataset grade. *Echocardiography* poses two questions on the same clip: a numeric task, the ejection fraction as a whole-number percentage scored against the reference value by MAE and R^2, and a categorical three-option item (reduced, mildly reduced, preserved) with shuffled letters, on which accuracy, balanced accuracy, macro F1 and recall of the abnormal categories are computed. *Melanoma relapse* is a two-option item (relapse within five years or not) posed on the whole-slide image together with the baseline metadata.

Evaluation protocol. DR: accuracy and balanced accuracy are computed on the three collapsed classes, QWK on the original five levels, and the referral-decision error rate counts predictions that cross the referral threshold in either direction; all DR metrics are unweighted means over the four cohorts. EF: categories are reduced (EF $\leq$ 40%), mildly reduced (40% $<$ EF $<$ 50%) and preserved (EF $\geq$ 50%) on the continuous reference; MAE is in EF percentage points. Relapse: recurrence F1 takes relapse as the positive class, PPV is precision among predicted relapse alerts, specificity is computed among non-recurrent patients, and the false-alert rate is the share of all evaluated cases receiving a false relapse alert. Baselines receive cine loops as temporally ordered frames at native resolution, sampled uniformly at 4 fps; whole-slide images, which commercial models cannot ingest at full resolution, are presented as the low-power overview that the agent's reader returns first, together with the clinical metadata. Confidence intervals are percentile 95% intervals from 1,000 case-level bootstrap replicates.

Held-out questions (Section 4.2). Three questions are asked of test images already in the benchmarks, and neither agent is trained on them; each agent receives the question with a one-sentence criterion and no skill text. *Diabetic macular edema*: the three-level IDRiD risk grade of the IDRiD images, which depends on the distance of hard exudates from the fovea. *Left-ventricular dilatation*: whether the reference end-diastolic volume of an EchoNet-Dynamic test loop exceeds 150 mL (Lang et al., 2015). *Ulceration*: presence of ulceration of the primary melanoma, scored against the VisioMel annotation, which is withheld from the inputs for this question. All three are scored by balanced accuracy (chance 33.3%, 50% and 50%).

C Examination skills and tool interface

Skill authoring. For each task, three board-certified subspecialists (5–9 years of experience, disjoint from the physicians who later evaluated the agent) drafted the examination skill independently and reconciled their drafts; the agreed document is the skill $\mathcal{S}$ used for demonstration, auditing, failure localization and the +Tools +Skill condition of Section 4.1.

Evidential principles. Seven principles were common to all skills: (i) coordinate discipline, with all spatial claims in original-image coordinates; (ii) landmark-derived crops, with every field of view positioned from measured anatomy rather than fixed grids; (iii) confirmation before acceptance, with no finding claimed without verification in the underlying data; (iv) instruments as proposal generators, with outputs adjudicated rather than adopted; (v) decision-directed code, with every execution justified by the pending clinical decision; (vi) coverage-bounded negative claims, with absence asserted only for anatomy actually inspected; and (vii) active disconfirmation, with competing hypotheses registered and tested before the conclusion.

Task workflows. *DR grading* (adaptive rather than gated): orient and survey in one compact pass (gradability, disc, fovea, arcades, laterality; posterior pole, four anatomy-oriented quadrants and visible periphery at a scale suitable for small red lesions) → resolve proliferative overrides first (NVD/NVE, preretinal or vitreous haemorrhage, fibrovascular proliferation, panretinal-photocoagulation pattern) with targeted original-coordinate crops viewed in native RGB and vessel-enhanced renderings → verify only the findings that determine the remaining grade boundary, adjudicating each candidate in native RGB (candidate pixels, blobs or connected components are proposals, never lesion counts) → when the moderate–severe boundary is relevant, execute the 4-2-1 rule in code from per-quadrant accepted haemorrhage counts, venous beading and IRMA, with tri-state arms for inadequately covered quadrants → assign the five-level grade; stopping rule: the fewest clinically sufficient operations. *EF categorization*: whole-clip read → interval segmentation proposing candidate cycles → phase-lock ED/ES against raw frames → final masks and disc-method EF in code → reconcile with visual impression → categorize; reference allocation of two frame-reader, three interpreter and two segmentation calls, which the agent may exceed when verification requires it. *Relapse prognosis*: case framing and priors → low-power map with adequacy check → medium-power survey of at least four distributed cells → high-power adjudication of at least three regions → coverage and pathology reconciliation separating morphology from covariates → calibrated risk decision.

Tools. *Code interpreter* (`code_interpreter`): stateful sandboxed Python with NumPy, OpenCV and PIL; each call submits the history of successful blocks followed by the new block, re-executes the history to restore state, runs the new block, and returns only artefacts created or modified in the current turn; failed blocks are not retained. *Frame reader* (`read_video`): returns temporally ordered frames over a time window, frame rate (or an automatic rate), resolution budget (small, normal or large) and frame cap that the policy sets in each call; every frame carries its exact timestamp and is saved to the workspace, where code opens it. CodeLancet uses all three resolution budgets, and about half of its calls read a narrow window around a candidate cardiac cycle after a whole-clip pass. *Segmentation tool* (`segmentation_tool`): a left-ventricular boundary model that, over an interval, returns per-frame chamber areas as a phase curve with candidate cycles ranked by depth and regularity of contraction; on a frame pair, returns masks, overlays and areas; computes no EF. *Whole-slide reader* (`read_wsi`): overview mode returns a low-power tissue footprint with a spatial grid and per-cell tissue scores; tile mode returns low-, medium- or high-power patches for the grid cells the policy names, at the corresponding pyramid level and a patch size it sets. The agent typically plans the cells in code from the overview, so the path through the slide is chosen per case. Neither reader performs any step of the examination: they are the equivalent of opening a file at a chosen scale. The segmentation tool was trained on EchoNet-Dynamic training data following the official setting; weights are released with the code. No learned perception model is used for fundus photographs at any stage: the fundus skill prescribes an adaptive examination carried out entirely in the code interpreter, and the fundus tool set consists of the code interpreter alone, for CodeLancet and for every frontier model in Section 4.1.

Execution semantics. The sandbox is a stateful Python workspace: variables, measured landmarks, coordinates and intermediate results persist across turns, so a crop position derived in one turn can be reused in the next, a measurement can be refined once new evidence arrives, and an instrument's output can be taken up in code and corrected, filtered or re-measured rather than accepted as returned. A failing code block returns its error message as an observation and is not retained, so the examination advances along a chain of successfully executed steps. The interpreter performs deterministic array and image operations and the instruments run with fixed weights, so the same code applied to the same image reproduces the same outputs; a saved trajectory can be re-executed from any step, and a

reviewer can amend one intermediate result, for example an end-systolic frame index or the accepted lesion candidates, and re-run what follows. Inference uses temperature 0.6, and the same sandbox service is used for trajectory generation, reinforcement learning and evaluation.

Fixed toolset (Section 4.2). The fixed-toolset variant replaces the interpreter by one tool per operation that the skills prescribe, each implemented with the same code the teacher's demonstrations execute: for fundus photographs, disc and fovea localization, quadrant cropping, native-resolution zoom, candidate detection, lesion verification, lesion counting and the 4-2-1 rule; for echocardiography, the frame reader, the segmentation tool, end-diastole and end-systole selection, mask-quality check, the method-of-discs EF and a report tool; for whole-slide images, the slide reader, tissue map, tile sampling, tile zoom, region scoring and risk aggregation. Teacher demonstrations are re-expressed as calls to these tools, and both training stages are run unchanged. The held-out questions would need exudate detection, exudate-to-fovea distance and the edema rule, a chamber-size tool and a threshold, and epidermis localization, ulcer detection and extent measurement, each as a further tool.

D TRAINING DETAILS

Demonstration audit (Section 2.1). The teacher (Qwen3.8-Flash-Next) samples up to eight examinations per training case at temperature 1.0 with the skill in its system prompt. Answer-correct rollouts are audited by a judge model that receives the skill, the full message history and every image the code produced, and returns whether each evidence-bearing step of the skill was executed with visible evidence, whether the conclusion follows from the recorded measurements, and whether any claimed check lacks a corresponding execution; self-reported checks without tool evidence do not count. A rollout is accepted when no step is missing or unsupported. The audit is adaptive: steps the skill makes conditional (for example the 4-2-1 count when proliferative disease has already been established) are not required.

Stage 1. SFT uses a learning rate of 10^{-5} for one epoch over $\mathcal{D}_0$, loss on assistant tokens only. GRPO uses a group size of $G = 8$, a global batch of 256 cases (2,048 examinations), three epochs, learning rate 10^{-6} with cosine decay to 5×10^{-7} and 5% warm-up, no KL regularization, asymmetric clipping ($\epsilon = 0.2$, upper 0.28), rollout temperature 1.0 with top-p 0.95, and at most 12 assistant turns of 4,096 tokens with 12 tool calls per examination.

Stage 2: localization and repair. The Stage-1 policy π_1 examines every training case once with real tool execution; messages, images, code history, tool outputs, sampled token ids and behaviour log-probabilities are stored. For failed examinations the judge (Gemini 3.6 Flash) receives the skill, the complete trajectory and its images, is told that it does not know the label and must not infer it, and returns either `abstain` or the id of the earliest assistant decision that the visible evidence shows to be problematic, with the skill requirement it violates. The state before that decision is restored by replaying the successful code history. From it the student samples four unguided continuations; correct ones are audited with the demonstration rubric, additionally checking that every new action is understandable from the public context alone. If none passes, the judge is asked for one next inspection action given only the public prefix and the skill; the hint must not state a finding, a count, a criterion outcome or an answer, is checked by a second action-only audit, and is placed in the system prompt for the first new turn only. Two guided continuations are sampled and audited in the same way, and the hint is deleted from the saved context, so training sequences never contain it. At most two suffixes are kept per root. Repair-SFT records are (public context, new assistant turn) pairs with cross-entropy on the new turn only; they are mixed with audited replays of π_1's correct examinations at a 7:3 ratio of actions and trained for two epochs at learning rate 10^{-5} with a global batch of 128.

Stage 2: Focal-PPO. Roots are the repaired roots, those with an accepted continuation, and, in a 7:3 ratio, initial states of training cases; roots left unrepaired are not used. From each root the frozen π_2 rolls out one unguided main rollout; its first $H{=}3$ assistant turns are the trained actions. Each non-terminal state before and after a trained action is valued by $K{=}4$ independent unguided continuations to termination under the state's remaining budget, $\hat{V}(s)$ being their mean reward; terminal states have value 0. The advantage of an action is $r_t + \hat{V}(s_{t+1}) - \hat{V}(s_t)$ with discount 1, shared by all its target tokens; continuations are never gradient records. The loss is

$-\frac{1}{|a|}\sum_{j\in a}\min\left(\rho_j\hat{A},\ \mathrm{clip}(\rho_j, 0.8, 1.28)\hat{A}\right)$ averaged over actions, that is, $\epsilon_{\mathrm{lo}} = 0.2$ and $\epsilon_{\mathrm{hi}} = 0.28$ in Eq. (2), where ρ_j is the token probability ratio against a frozen copy of π_2 whose log-probabilities are recomputed and checked against the sampling engine before the first update. Zero-advantage actions stay in the batch. Training runs three epochs over the fixed action pool at a global batch of 256, learning rate 10^{-6} with cosine decay to 5×10^{-7} and 5% warm-up, weight decay 0.1, gradient clipping 1.0, bf16, no KL term. Sampling for main rollouts and continuations uses temperature 1.0 and top-p 1.0. The second-GRPO-stage alternative in Table 3 starts from the same π_2 with the Stage 1 GRPO configuration; the initial-state-only alternative uses the same estimator, budget and number of roots, drawn entirely from initial states.

Random intermediate roots. To separate the localized decision from restarting inside a failed examination at all, each repair root is replaced by an intermediate state of the same failed examination at a random depth, with depths matched to the distribution of the repair roots; the estimator, budget, number of roots and 7:3 mixture are unchanged. These roots trail repair roots on every task (Figure 9). Restarting inside failed examinations lowers the defect rate from 19.8% (initial states only) to 15.9% but leaves accuracy near the initial-state level; the accuracy gain and the rest of the defect reduction, to 13.1%, come from where the judge localizes the failure.

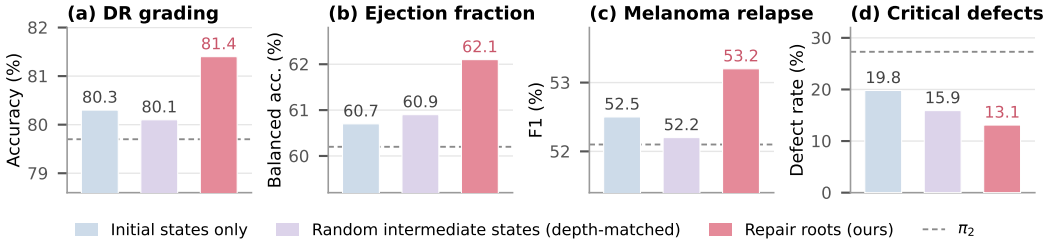

Figure 9: **Root placement for Focal-PPO.** Same estimator and budget from π_2 (dashed); roots at initial states, at depth-matched random intermediate states of failed examinations, or at repair roots.

Remark (outcome credit versus decision credit). For a group of G rollouts with coinciding terminal rewards, $r_1 = \cdots = r_G$, the numerator of the group-relative advantage $(r_i - \bar{r})/\sigma_r$ vanishes and the group is dropped ($\sigma_r = 0$), so no decision of any rollout receives a gradient, whatever those decisions were. With independent continuations, the Focal-PPO estimate $\hat{A} = r_t + \hat{V}(s_{t+1}) - \hat{V}(s_t)$ satisfies $\mathbb{E}[\hat{A} \mid s_t, a_t] = A^{\pi_2}(s_t, a_t)$, the change in expected reward caused by a_t, which is non-zero whenever the decision matters, even when every sampled rollout of the case ends alike; a finite-K estimate can still be zero or non-zero by chance. Outcome credit is therefore silent on exactly the cases a converged policy mostly produces; decision credit is not. Focal-PPO places its roots where a non-zero estimate is more likely: from a repaired root, the examination has been observed both to fail and to be completed correctly.

Judge audit of test examinations. The defect rates in Table 3 and Figure 7c apply the demonstration rubric, with the same judge prompt and without the label, to every test examination of each checkpoint; an examination is counted as defective when the judge returns at least one critical missing or unsupported step.

E BASELINE PROMPTS

Baselines received the same clinical question and answer space as the agent, without tools. For DR the question was a five-option multiple-choice item, with minor dataset-specific formatting differences:

```
Based on this fundus image, what is the severity level of
diabetic retinopathy?
A. No DR
B. Mild non-proliferative DR
C. Moderate non-proliferative DR
```

```
D. Severe non-proliferative DR
E. Proliferative DR
```

Models that did not return a bare option letter were additionally instructed to reply with the letter only. For numeric EF estimation:

```
The following {N} images are frames sampled uniformly (in
temporal order) from an apical-4-chamber echocardiogram
video of a single patient.  Estimate the left ventricular
ejection fraction (LVEF) in this echocardiogram as a
whole-number percentage.  Do not return an EF category or a
multiple-choice letter.  Return only one compact JSON object
of the form {"lvef":N}, where N is a base-10 integer from 0
to 100 with no percent sign.
```

The categorical EF and relapse questions were presented as multiple-choice items over the option sets defined in Section 3.1.

F FULL BENCHMARK RESULTS

Tables 8 and 9 give the complete metric panels for the three tasks, and Figures 10 and 11 plot every DR and EF metric with bootstrap intervals.

Table 8: **Ejection-fraction assessment: complete metrics** on the EchoNet-Dynamic test partition ($n = 1{,}277$). Abn. F1: F1 for abnormal (mildly reduced or reduced) LVEF. PCC/CCC: Pearson and concordance correlation of the numeric estimate with the reference.

Model	Acc	Bal. acc	Macro F1	Mild rec.	Red. rec.	Abn. F1	MAE↓	RMSE↓	R^2	PCC	CCC
GPT-4o	22.6	34.0	20.9	27.2	58.8	34.8	10.47	14.2	−0.33	0.03	0.03
GPT-5.4 Mini	48.4	34.3	32.1	28.0	19.4	31.0	13.59	16.9	−0.92	−0.02	−0.02
Gemini 2.5 Flash	60.5	34.3	34.1	11.2	18.1	29.1	12.98	17.7	−1.10	0.09	0.09
Claude Sonnet 4.6	68.0	32.3	30.7	0.0	11.3	13.9	12.72	14.8	−0.47	−0.01	0.00
GPT-5.6 Luna	66.2	33.0	32.9	3.2	13.1	21.1	10.66	13.6	−0.24	0.04	0.03
Gemini 3.5 Flash Lite	69.5	37.2	35.0	0.0	26.3	27.0	12.14	16.9	−0.91	0.09	0.09
CodeLancet (ours)	**69.7**	**62.1**	**54.6**	**40.8**	**72.5**	**58.3**	**7.32**	**9.6**	**0.38**	**0.67**	**0.65**

Table 9: **Diabetic retinopathy grading and melanoma relapse prognosis: complete metrics.** DR values are unweighted means over the four cohorts; ordinal MAE is in severity levels. Relapse metrics on VisioMel ($n = 1{,}322$).

	DR grading					**Melanoma relapse**					
Model	Acc	Bal. acc	QWK	Ord. MAE↓	Ref. err↓	Acc	F1	Spec	PPV	False alert↓	Total err↓
GPT-4o	59.6	52.9	45.1	0.84	27.2	78.4	50.9	80.0	40.0	16.8	21.6
GPT-5.4 Mini	61.9	49.1	41.9	0.84	30.8	75.0	47.7	75.7	35.9	20.4	25.0
Gemini 2.5 Flash	55.5	40.5	20.9	1.04	36.9	66.6	44.2	63.5	30.2	30.6	33.4
Claude Sonnet 4.6	63.3	49.7	46.1	0.84	25.5	67.7	45.7	64.4	31.3	29.9	32.3
GPT-5.6 Luna	71.3	57.1	55.3	0.69	21.6	60.7	34.2	60.2	23.4	33.4	39.3
Gemini 3.5 Flash Lite	75.2	54.7	57.9	0.65	17.8	60.0	35.7	58.2	24.1	35.1	40.0
CodeLancet (ours)	**81.4**	**61.5**	**74.9**	**0.47**	**9.5**	**81.6**	**53.2**	**84.8**	**45.0**	**12.8**	**18.4**

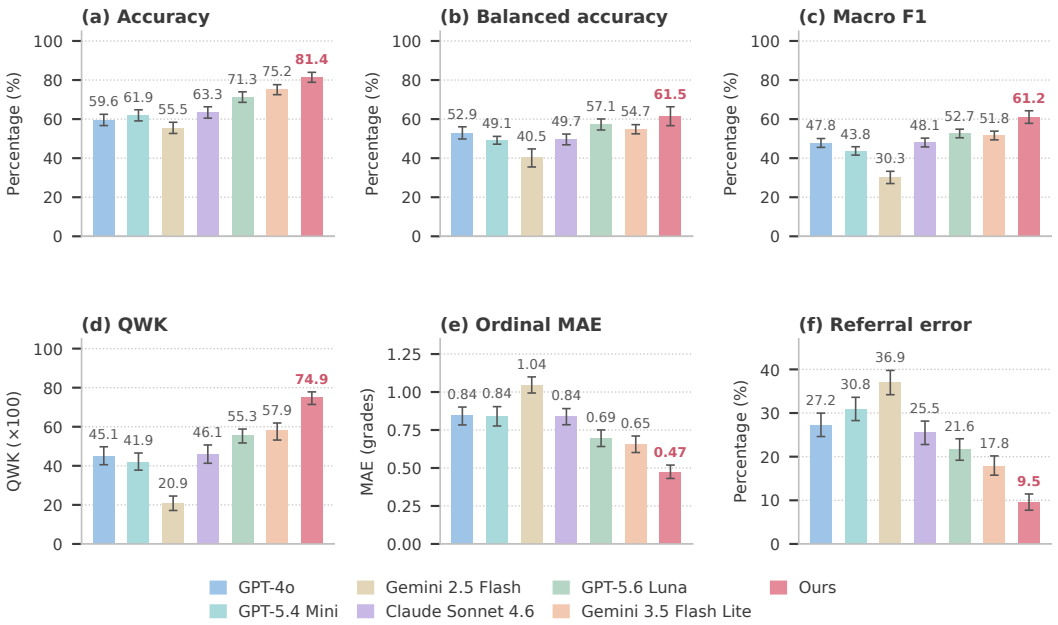

Figure 10: **DR grading: all metrics with 95% bootstrap intervals** (unweighted means over the four cohorts).

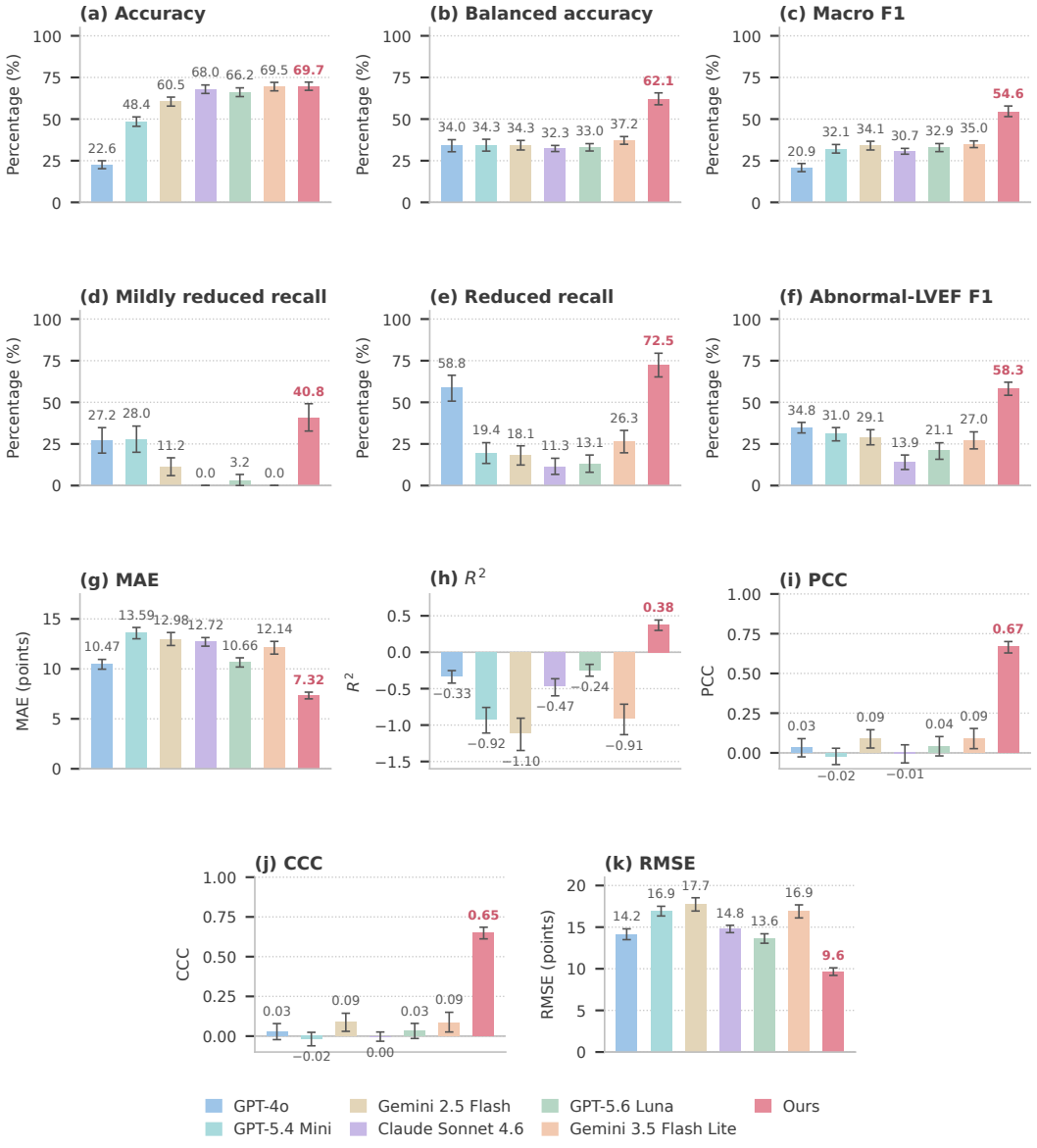

Figure 11: **EF assessment: all metrics with 95% bootstrap intervals.**

G ATTRIBUTION STUDY: PROCESS STATISTICS

The skill prompt for each task was the same document used to train CodeLancet (Appendix C), appended verbatim after the base system prompt and tool schemas. Metrics and case sets are those of Table 1. Table 10 gives tool-call counts and sequence adherence for each frontier model and condition.

Table 10: **Process statistics for frontier models with tools, without and with the skill prompt.** Mean tool calls per case and, for the skill condition, fraction of cases that followed the skill's prescribed tool sequence exactly (not defined for DR, whose skill uses the interpreter alone). DR values are pooled over the four cohorts.

		+Tools	**+Tools +Skill**	
Task	Model	Tool calls	Tool calls	Sequence adh. (%)
DR grading	GPT-5.6 Luna	2.2	3.1	–
	Gemini 3.5 Flash Lite	2.8	2.7	–
	Claude Sonnet 4.6	6.1	6.7	–
EF categorization	GPT-5.6 Luna	2.1	7.0	61.9
	Gemini 3.5 Flash Lite	2.8	6.5	12.5
	Claude Sonnet 4.6	4.0	7.0	94.8
EF estimation	GPT-5.6 Luna	2.5	7.0	66.6
	Gemini 3.5 Flash Lite	3.3	6.4	22.8
	Claude Sonnet 4.6	4.7	7.0	89.8
Relapse prognosis	GPT-5.6 Luna	3.0	5.2	99.7
	Gemini 3.5 Flash Lite	2.9	5.2	19.4
	Claude Sonnet 4.6	4.9	5.5	98.7

H INSTRUMENT ABLATION, ADJUDICATION AND CORRECTION

Table 11 breaks down the faulty segmentation proposals of Table 6 by failure mode and by the agent's response. In the code-only conditions of Table 5, the interpreter receives the removed segmentation allowance, so the total tool budget is unchanged. On fundus photographs, where the code interpreter is the only tool, the agent's code localizes the disc or fovea in 87% of cases, partitions the image by anatomical quadrant in 79%, builds classical lesion detectors in 51%, and verifies candidate findings in native-resolution crops in almost every case.

Table 11: **Taxonomy of faulty instrument output and agent response.** Breakdown of the faulty segmentation proposals in Table 6 by failure mode and by how CodeLancet responded.

Task	Failure mode	Share (%)	Rejected & re-run (%)	Overridden in code (%)	Adopted (%)
EF	Mask leaks into atrium / outflow	47.0	33.7	38.4	27.9
	Wrong ED/ES frame locked	36.1	36.4	36.4	27.3
	Under-segmentation (apex or lateral wall)	16.9	19.4	32.3	48.4

Faults that the raw frames can contradict directly (a wrong cardiac phase, a mask that visibly crosses the mitral plane) are the ones the agent most often catches, and more than half of the faults it detects are corrected in its own code rather than by re-running the instrument.

Within Stage 1, the reward signal of outcome GRPO is concentrated on the cases with a faulty segmentation proposal (Table 6): they constituted 14.6% of the echocardiography training cases but 41.3% of the cases retained after zero-advantage filtering, averaged over training, and their mean within-group reward standard deviation was 0.39 against 0.11 for cases with a sound proposal.

I PHYSICIAN EVALUATION PROTOCOL

Sample and reviewers. 150 trajectories of CodeLancet were reviewed: 50 fundus cases sampled uniformly across severity grades; 50 echocardiography cases (30 categorical, 20 numeric); 50 pathology cases balanced between relapse and non-relapse. Three board-certified subspecialists per modality (6–11 years of experience), disjoint from the skill authors, served as independent reviewers after a joint calibration session.

Design and rubric. Reviewers received each trajectory in full (every deliberation, every executed action with its returned evidence, the final answer) in randomized order, blinded to the reference label and to correctness. Case-level dimensions: *evidence grounding* (findings and conclusion supported by the images and tool results) and *expert-workflow concordance* (the examination addressed the question through an appropriate sequence of observations and checks). Step-level dimensions: the *necessity* of each investigative action and the *correctness* of its interpretation. Items were scored 0–5 (5 fully meets, 3 partly meets, 0 does not meet); endorsement denotes a score $\geq$4. Reviewers used a web interface presenting the source media, each action with its returned evidence, and the rubric. Endorsement rates per modality and per dimension, and inter-rater agreement as the mean pairwise quadratic-weighted kappa among the three raters, are reported in Table 7.