SONOARENA: FORMULATING AND BENCHMARKING AGENTIC DECISION-MAKING IN ULTRASOUND EXAMINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultrasound examination is a sequential decision-making process: a sonographer interprets the evidence acquired so far, decides where to scan next, and stops once the evidence is sufficient. Previous work on ultrasound AI has largely overlooked this decision loop. In this work, we formulate ultrasound examination as a partially observable Markov decision process (POMDP), in which diagnostic evidence is revealed only when the corresponding acquisition action is taken. Building on this formulation, we introduce SONOARENA, an interactive benchmark in which ultrasound agents sequentially acquire diagnostic evidence and decide what to report. We theoretically characterize the operating regimes of different policy types, which is then supported by the empirical results. We benchmark 24 Large Language Model (LLM)-based policy configurations, spanning general-purpose and task-trained policies. The strongest task-trained policy outperforms the evaluated baselines overall, but remains below the reference, leaving room for more targeted training methods. These results demonstrate the necessity and feasibility of evaluating and training ultrasound agents not only by how well they interpret acquired observations, but also by whether they can make the right decision about what to acquire to support their conclusions. Code is in supplementary material.

1 INTRODUCTION

Ultrasound is among the most widely used imaging modalities, yet conducting an effective examination requires expertise that is often scarce, particularly in primary care, emergency medicine, and low-resource settings (Andersen et al., 2019; Abrokwa et al., 2022). Unlike radiography, CT, and MRI, where acquisition typically precedes interpretation, ultrasound interleaves the two. In ultrasound, each image is an examiner-selected view whose diagnostic content depends on probe position and orientation. The examiner interprets each view in real time, uses the evidence to select the next acquisition, and stops when sufficient evidence has been collected. What can be diagnosed from an examination therefore varies according to what has been acquired, while what should be acquired depends on what has already been observed (van Sloun, 2024). Ultrasound examination is thus not simply the interpretation of a given input, but a closed loop between evidence acquisition and diagnostic reasoning (Figure 1).

Despite this closed-loop nature, existing ultrasound AI captures only parts of the examination loop. Vision models interpret pre-acquired images or videos (Le et al., 2026; Weng et al., 2025; Ge et al., 2025; Azhar et al., 2026) while providing limited support for acquisition decisions. Guidance and robotic systems pursue predefined views (Yue et al., 2025; Ao et al., 2025; Jiang et al., 2025; Bi et al., 2026) while overlooking the adaptive nature of acquisition which may affect diagnostic effectiveness. Beyond ultrasound, interactive medical agents gather evidence through follow-up questions, diagnostic tests, or record navigation (Li et al., 2024; Nori et al., 2025; Gao et al., 2026; Wu et al., 2026; Qiu et al., 2025), but they operate over high-level clinical decisions that require little physical interaction. Taken together, prior work does not address the within-modality policy for ultrasound examination: selecting acquisition actions, deciding when to stop, and ensuring that each report claim is sufficiently supported. In ultrasound scanning, absence of evidence does not imply evidence of absence. An unobserved finding may be truly absent or merely missed, while adequate acquisition supports reporting absence and inadequate acquisition calls for reacquisition. Static evaluations fail to capture this dynamic nature of ultrasound examination. Detailed related work is deferred to Appendix A.

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.

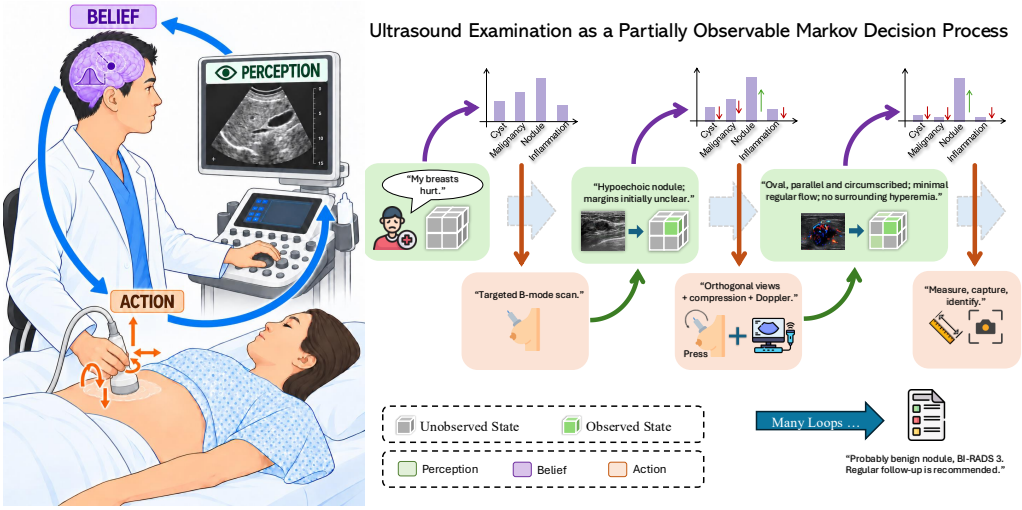


Figure 1: **Ultrasound examination as a POMDP.** The examiner alternates among perception of the current view, a belief over findings not yet observed, and the next acquisition action (left). Each scan reveals only part of the latent state; subsequent actions are chosen from the updated belief, and the loop repeats until the evidence licenses a report (right).

To address this gap, we propose SONOARENA, an interactive benchmark that turns static ultrasound reports into a dynamic environment. Given an indication and a limited budget, an agent can acquire action-dependent evidence, decide when to stop, and produce a diagnostic report within the environment. A post-hoc scoring pipeline evaluates the acquisition trajectory, verdict, and report. Formally, we model ultrasound examination as a partially observable Markov decision process (POMDP), whose latent state captures patient findings and acoustic-window adequacy. The formulation enables us to theoretically characterize the conditions for each type of policy to work more effectively: when a typical patient state dominates the prior and acquisitions are reliable, replaying a fixed protocol has bounded loss; when an observation changes whether a later acquisition is worthwhile, every predefined policy can incur non-zero loss. Empirical evidence supports these analyses.

We benchmark 24 LLM-based policies, spanning general-purpose and task-trained policies, alongside programmatic reference policies. Performance varies across evidence acquisition, structured verdicts, and free-text faithfulness. General-purpose LLM-based policies perform better in settings that require redirection, but sometimes trail simple programmatic policies when the scanning protocol is stable. Task-trained policies show the reverse pattern: supervised fine-tuning (SFT) learns protocol following and reacquisition, while group relative policy optimization (GRPO) recovers adaptation. This variation across examination settings matches our theoretical analysis. Performance generally improves with backbone model scale, and no policy leads in every regime. Interestingly, better acquisition does not always yield a more faithful report. Several agents that collect more evidence also add more unsupported claims in free text. Task-trained policies make fewer such errors despite having no dedicated objective for report faithfulness.

Together, our contributions are threefold:

1. **Problem formulation.** We formulate ultrasound examination as a POMDP that separates patient findings from acoustic-window adequacy, and characterize when fixed protocols suffice and when observation-conditioned acquisition is required.
2. **SONOARENA benchmark.** We introduce a closed-loop benchmark with three report-grounded families and a synthetic stress test, providing action-dependent observations, counterfactual trajectories, and separate evaluation of acquisition, verdicts, and reports.
3. **Policy diagnosis.** Across 24 LLM-based policies and programmatic references, we show that policy rankings vary across acquisition regimes and evaluation axes, exposing complementary failures in general-purpose and task-trained policies.

2 SONOARENA: ULTRASOUND EXAMINATION AS A POMDP

In this section, we formalize ultrasound examination as sequential decision-making under partial observability, then describe how SONOARENA instantiates and evaluates that process. Figure 2

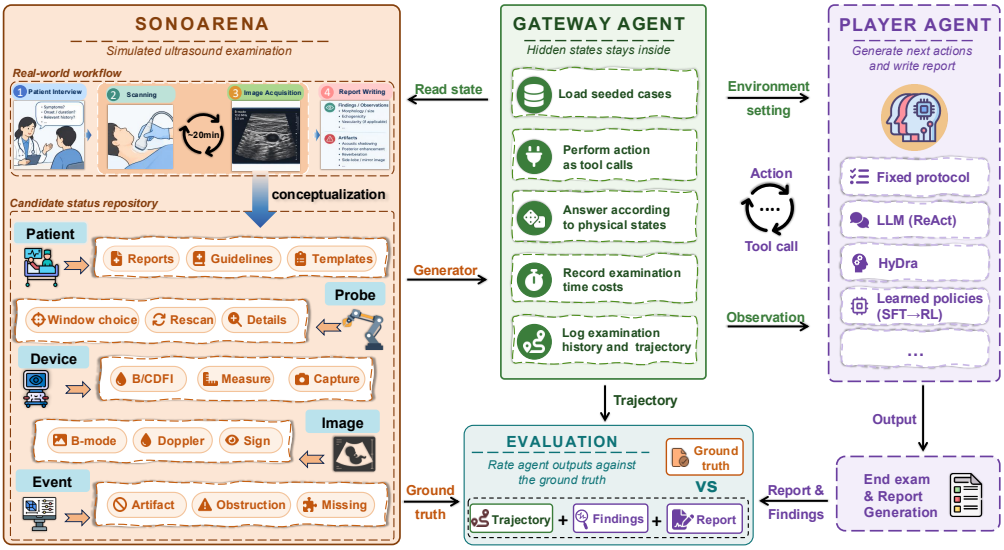


Figure 2: **SonoArena overview.** The environment separates patient state, device controls, probe actions, ultrasound observations, and contingency events. A gateway keeps the state hidden while executing the examination agent's tool calls, metering time, and recording the episode. After termination, the scorer evaluates the trajectory, structured findings, and report against ground truth.

summarizes the interaction: a player requests scans through a gateway, receives the resulting evidence, and eventually submits findings and a report. Section 3 then develops a series of theoretical analyses of player policies.

2.1 Ultrasound Examination as a POMDP

A partially observable Markov decision process (POMDP) models sequential decision-making when agents can reveal only parts of a hidden environment through their actions and must infer its state from the resulting observations (Kaelbling et al., 1998). It is specified by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, T, Z, R, b_0)$, consisting of state, action, and observation spaces; transition and observation models; a reward; and an initial belief. The interaction history $h_t = (a_1, o_1, \ldots, a_t, o_t)$ induces a posterior belief $b_t(s) = \Pr(s \mid h_t)$, which can be used for choosing the next action.

Ultrasound has precisely this structure. The pathology is hidden, and even a present finding may not be visible from every probe position. Each acquisition therefore reveals only a window-dependent, imperfect view; and its result may change where to scan next or what work-up is required. Additional views can improve the report, but also consume the limited examination time.

Formally, the hidden patient state can be denoted by $s = (y, \Omega)$, where y records which findings are present, and Ω records whether each finding is visible through each acoustic window. The action space $\mathcal{A}$ contains probe, device, work-up, and terminal reporting actions. Each non-terminal action returns an observation o_t, which accumulates in h_t.

We assume that pathology and acoustic access do not change during an examination, therefore the transition model is the identity, $T(s' \mid s, a) = \mathbf{1}[s' = s]$. Uncertainty lies in the observation model $Z(o \mid s, a)$. For a finding anatomically reachable from window w, we use

$$\Pr(\text{sign} \mid y_i{=}1, \Omega_{wi}{=}1) = 1 - \beta_i, \quad \Pr(\text{sign} \mid y_i{=}0, \Omega_{wi}{=}1) = 0, \quad \Pr(\text{sign} \mid y_i, \Omega_{wi}{=}0) = 0.$$

The parameter β_i describes a stochastic miss that rescanning may repair. When $\Omega_{wi} = 0$, the finding remains invisible from window w and must be sought from another window. Thus, failure to observe a sign can mean either that the finding is absent or that the acquisition was unsuccessful.

The terminal return trades report utility against acquisition cost:

$$R(s, \tau, \hat{r}) = \sum_{i=1}^{m} u(\hat{r}_i, y_i, \Omega_{\cdot i}, h_T) - \lambda \sum_{a \in \tau} c(a). \tag{1}$$

Here, $\hat{r}_i$ is the verdict for finding i, and h_T is the final history with action sequence τ. The utility u evaluates each verdict against the hidden finding and the evidence acquired. A negative verdict is rewarded only if h_T contains an adequate view that could have revealed the finding (Appendices B.2 and E). Finally, the clinical indication sets the initial belief $b_0(y, \Omega)$ over pathology and acoustic access, with subsequent observations updating that belief throughout the examination.

2.2 Five Components of the Ultrasound World

SonoArena implements the model as a seeded, interactive ultrasound environment. Its five components define the patient state, the action space and the possible observations.

Patient state (y, Ω)**.** The variable $y_i \in \{0, 1\}$ indicates whether finding i is present. Let the family-level map $A_{wi} \in \{0, 1\}$ indicate whether window w anatomically reaches finding i, and the patient-specific variable $\omega_w \in \{0, 1\}$ indicate whether that window is accessible. Their product, $\Omega_{wi} = A_{wi}\omega_w$, determines whether finding i can be revealed through w. In ultrasound examination, for example, bowel gas may set $\omega_w = 0$ even when $A_{wi} = 1$.

Scanning actions. Scanning is the primary means of acquiring diagnostic evidence in ultrasound. Automated scanning methods have focused on probe control and image quality (Yue et al., 2025; Jiang et al., 2025; Bi et al., 2026). We abstract these low-level controls and evaluate whether the examination has acquired enough diagnostic information. The action $\textsc{scan}(w)$ places the probe at window w and returns a survey view, which may reveal partial evidence. The player can repeat the scan or move to another window when the current one is inadequate. These are the examination-level choices to "keep scanning" or "look elsewhere."

Ultrasound device actions. Device controls determine what anatomical information is acquired and which views are retained as evidence. The device maintains an imaging mode $\mu \in \{\text{B-mode}, \text{CDFI}\}$. The action $\textsc{mode}(\mu)$ changes this mode, while $\textsc{capture}(w)$ stores an acquired view as evidence for the final report.

Ultrasound observations. Ultrasound image interpretation is well studied (Le et al., 2026; Ge et al., 2025; Weng et al., 2025). To separate visual perception from planning, we give every player the same natural-language interpretation of each acquired image. The player decides whether the resulting evidence is diagnostically adequate and uses it to write the report. It may also call $\textsc{detail}(w)$ for morphological features or $\textsc{measure}(w)$ for measurements.

Contingency events. At the start of each episode, the environment samples persistent access obstacles into ω and then holds them fixed. The observation model Z samples retryable misses and transient acquisition or device failures when an action is executed. If the source report does not specify a counterfactual outcome, declared and auditable rules supply it and such content is not treated as observed ground truth. A shared seed governs all stochastic events, so repeated actions remain meaningful and players can be compared under paired conditions.

2.3 Gateway and Evaluation

The gateway mediates all interaction between the player and the arena. At reset, it loads a fixed case and seed while keeping the hidden state (y, Ω) inside the environment. The player observes only the indication, available tools, remaining budget, and action–observation history. For each requested action, the gateway executes the corresponding tool call, applies T and Z, returns the observation, charges $cost(a)$, and records the transition in the episode log. The episode ends when the player submits a report or exhausts the budget.

Evaluation occurs after the episode and cannot influence the examination. The scorer compares the hidden ground truth with three artifacts produced by the player: the acquisition trajectory, structured findings, and the free-text report. The trajectory records coverage, usable views, and acquisition time. Structured findings determine the episode return R (Eq. 1) and whether negative claims are licensed by the acquired evidence. An LLM attributor parses the report into claims, which are then checked by deterministic rules for unsupported or inconsistent content. All adjudication after attribution is deterministic (Appendix B.7).

3 PLAYER POLICIES: LIMITS, REFERENCES, AND LEARNING

We investigate three aspects of player policies: (i) when adaptation is necessary, (ii) what can serve as a reference when the optimal policy is intractable, and (iii) whether a language model can learn such a policy.

3.1 WHEN DOES ADAPTIVITY MATTER?

We begin by asking whether acquisition decisions need to depend on observations. A *fixed protocol* follows a predefined procedure, while an *adaptive policy* conditions its decisions on the interaction history. We show that when a typical patient state dominates the prior, a fixed protocol is near optimal. Let V^π denote its expected return, $V^\star$ the optimal value, and Δ the return range.

Proposition 1 (Predictable examinations need little adaptation). *Suppose one state $s^\star$ has prior probability at least $1-\delta$, and its optimal noiseless trajectory contains N observations, each corrupted with probability at most β. Replaying that trajectory as a fixed protocol π_{OL} gives*

$$V^\star - V^{\pi_{\mathrm{OL}}} \le (\delta + N\beta)\Delta.$$

When the typical state has high prior probability and acquisitions are reliable, small δ and β yield a small return gap (proof in Appendix B.3). This regime resembles screening ultrasound, where clinicians apply a standardized protocol to a relatively predictable population.

Fixed protocols become insufficient when an observation changes the value of a later acquisition. Consider a cue observed at w_0, after which the policy may acquire w'. Let q denote the probability mass of states in which w' reveals a present finding, p the mass in which its k' findings are absent, and e the remaining mass. Let g and g^- denote the corresponding gains over abstention.

Proposition 2 (Actionable observations require adaptation). *If w_0 reaches neither the target finding nor those assessed at w', and prepending w_0 does not lower open-loop value, then every open-loop policy π satisfies*

$$V^\star - V^\pi \ge \min\{p(\lambda c(w') - k'g^-), q(g - \lambda c(w'))\} - e\Delta.$$

Observation error at most β reduces this bound by at most $\beta\Delta$.

A fixed protocol must either always acquire w' or always skip it. Always acquiring w' incurs excess cost $\lambda c(w') - k'g^-$ on the p states; always skipping it forfeits net diagnostic gain $g - \lambda c(w')$ on the q states. An adaptive policy in contrast can follow the cue. This occurs in diagnostic examinations, where a finding or a failed window may change the next action (proof in Appendix B.4).

3.2 TWO REFERENCES: CLAIRVOYANCE AND HYDRA

Evaluating and training adaptive policies requires tractable references. Since computing $V^\star$ is generally intractable, we use a *clairvoyant* policy as a performance ceiling and a custom Hypothesis Driven Agent (HYDRA) as an executable teacher.

The *clairvoyant* has access to the hidden state (y, Ω) and selects the best route for that state. This yields a prize-collecting weighted set-cover problem, which we solve exactly by exhaustively enumerating all window subsets. We denote its maximum return by $\bar{V}$, then we have

$$V^\pi \le V^\star \le \bar{V}, \qquad \bar{V} - V^\pi = \underbrace{\bar{V} - V^\star}_{\text{price of clairvoyance}} + \underbrace{V^\star - V^\pi}_{\text{policy gap}}.$$

The clairvoyant policy observes the hidden state and therefore provides a performance ceiling rather than realistic examination behavior. HYDRA receives no hidden-state information. It instead maintains a belief b using the observed history and the case generator. Let $U^*(b)$ denote the maximum expected utility of reporting immediately under belief b. For each action a, HYDRA considers plans of up to M repeated looks that stop at the first observed sign. Let b_m^a denote the resulting belief under a plan of m looks. The expected value of information is

$$\mathrm{EVOI}(a) = \max_{1 \le m \le M} \{\mathbb{E}[U^*(b_m^a)] - U^*(b) - \lambda m\, c(a)\}.$$

At each step, HYDRA executes the action with the largest positive EVOI, updates its belief, and replans. It stops when no acquisition has positive EVOI or the budget is exhausted. We use $M = 4$ in our experiments.

When $M = 1$, this decision rule is Bayes-optimal under one-step lookahead by construction. Such local optimality does not generally imply global optimality. Several acquisitions that are individually unprofitable may become valuable together, for example by crossing an evidence threshold or revealing an accessible route across windows. Proposition 3 identifies a canonical repeated-acquisition setting in which the one-step rule is nevertheless globally optimal.

Consider one finding and one known-open window that can be scanned repeatedly. Let b be the prior probability that the finding is present, β the miss probability, c the cost of one look, $G = u_{\text{TP}} - u_{\text{FN}}$, and $g = u_{\text{TP}} - u^{+}_{\text{IND}}$. Let $p^{\dagger}$ denote the reporting threshold.

Proposition 3 (Canonical repeated acquisition). *For a policy that takes $n \geq 1$ looks and reports* ABSENT*,*

$$\bar{V} - V(n, \text{ABSENT}) = b\left[\beta^n G + \lambda c\left(\frac{1-\beta^n}{1-\beta} - 1\right)\right] + (1-b)(n-1)\lambda c.$$

Moreover, if $p^{\dagger}(1-\beta)g > \lambda c$, then HYDRA *with $M = 1$ attains $V^{\star}$.*

The condition requires the expected gain from one more look to exceed its cost. Under this condition, HYDRA attains the global optimum even when the examination requires several looks. This guarantee does not extend to the general multi-finding, multi-window setting, where threshold crossing, uncertain access, and complementary acquisitions can make several acquisitions jointly valuable. Appendix B.11 describes the effect of $M > 1$ and the detailed cost approximation used.

3.3 Learning from Verifier Feedback

HYDRA depends on the simulator probabilities to maintain an explicit belief model. To obtain a policy that acts from the interaction history without this belief model, we train Qwen3-8B (Yang et al., 2025) in two stages. SFT imitates reference trajectories, and GRPO (Shao et al., 2024) subsequently optimizes the policy using verifier feedback.

For each case, we generate stochastic trajectories with HYDRA and a programmatic policy, retain the trajectory with the higher return, and train the model using response-masked cross-entropy. We then apply GRPO to rollouts sampled from the SFT policy. All rollouts in a group share the same case s and simulator seed. For rollout j, the verifier computes

$$R_{sj} = R(s, \tau_j, \hat{r}_j), \qquad A_{sj} = \frac{R_{sj} - \mu_s^R}{\sigma_s^R + 10^{-4}},$$

where μ_s^R and σ_s^R are the within-group mean and standard deviation of the return. The episode-level advantage A_{sj} is assigned to every decision turn. Sharing the case and simulator seed makes the update compare alternative examinations under paired conditions. Full training details are in Appendix B.8.

4 Experimental Setup

Datasets. We evaluate four ultrasound examination families: thyroid, breast, liver, and focused emergency. The first three use the official train/test splits of three public structured-report datasets (Li et al., 2025), comprising $1{,}399/412$, $1{,}256/395$, and $498/168$ cases, respectively. For these families, the source reports define the hidden findings, while clinical indications and acoustic-access events are sampled once per random seed. Focused emergency is a synthetic stress test with $60/60$ train/test cases. Appendix B.6 describes the construction process.

Policies. All evaluated agents receive the same indication, tools, observations, and budget. We evaluate three groups: (i) *programmatic policies*, including a fixed protocol and a stronger variant that reacquires unusable views and abstains from unsupported negative findings; (ii) *general-purpose LLM-based policies*, implemented by placing each general-purpose LLM in the same ReAct agent harness to select actions, observe the results, and decide whether to continue or produce the final structured findings and report (models listed in Table 1, harness details in Appendix B.9); and (iii) *task-trained LLM-based policies* using `Qwen3-8B`. We evaluate the base model, SFT, GRPO applied directly to the base model, and SFT→GRPO. Section 3 describes the training methods, with

Table 1: **Results of different policies on SONOARENA.** Green (1st), purple (2nd), and yellow (3rd) backgrounds are to distinguish the top three results. Darker blue indicates better performance.

Policies	SonoScore by exam. family↑ (based on trajectory)				Overall SonoScore ↑	Evidence (based on finding)			
	Thyroid	Breast	Liver	Focused em.		Cov ↑	U.N. ↓	U.P. ↓	Abst. ↓
Privileged References									
Clairvoyant	78.4±0.2	83.6±3.1	99.0±0.4	74.2±3.2	83.8±1.5	86.4±0.6	0.00±0.00	0.00±0.00	17.4±0.7
HYDRA-Teacher	59.4±1.4	59.2±3.3	94.8±3.0	53.9±0.7	66.8±0.6	89.9±0.5	0.00±0.00	0.00±0.00	24.4±0.6
Programmatic Policies									
Fixed protocol	18.4±4.0	29.7±2.9	57.0±2.3	-2.1±5.9	25.8±2.3	70.3±0.9	1.15±0.04	0.00±0.00	16.4±0.2
+ re-look + abstain	53.6±3.7	62.4±1.9	65.2±0.3	4.8±1.0	46.5±0.8	78.3±0.8	0.00±0.00	0.00±0.00	63.0±0.3
General-Purpose LLM Policies									
`gpt-4o-mini`	11.2±4.1	21.5±3.8	41.6±2.5	20.6±5.3	23.7±2.0	24.6±0.9	0.01±0.01	0.08±0.02	75.8±0.6
`gpt-5.4-mini`	22.4±4.9	30.7±4.6	51.4±3.5	40.9±6.7	36.4±2.5	35.4±1.2	0.13±0.04	0.06±0.02	63.2±0.9
`gpt-5.5`	36.2±8.6	58.9±4.9	79.3±3.2	46.1±2.1	55.1±2.8	79.8±0.8	0.00±0.00	0.00±0.00	23.3±0.5
`claude-sonnet-4-6`	21.7±5.9	37.5±5.7	68.2±6.9	57.4±5.6	46.2±3.0	59.5±1.8	0.13±0.03	0.02±0.01	40.9±1.2
`claude-sonnet-5`	25.4±4.9	38.0±5.7	63.2±7.0	49.6±6.3	44.1±3.0	66.4±1.6	0.13±0.04	0.04±0.02	34.4±1.1
`claude-opus-5`	29.5±5.6	44.4±5.5	72.5±7.1	64.2±5.7	52.7±3.0	63.9±1.7	0.03±0.02	0.02±0.01	38.3±1.1
`gemini-3.1-pro`	35.1±8.1	46.3±8.9	75.9±1.7	49.8±7.8	51.8±4.7	74.6±3.1	0.01±0.00	0.01±0.00	27.7±2.7
`gemini-3.8-flash`	20.1±6.2	55.2±6.3	77.9±8.0	54.6±6.0	52.0±3.3	76.5±1.4	0.02±0.01	0.01±0.01	26.3±1.0
`grok-4.6`	36.3±11.1	58.3±13.2	57.1±16.7	53.7±9.0	51.4±6.4	70.2±2.8	0.00±0.00	0.04±0.03	29.6±1.8
`glm-5.3`	48.4±10.1	34.8±8.8	61.6±15.0	52.8±10.9	49.4±5.7	73.1±2.8	0.02±0.02	0.01±0.01	28.9±2.0
`kimi-k3`	6.1±3.1	9.3±4.1	16.4±4.8	23.0±6.4	13.7±2.4	21.2±2.5	0.01±0.01	0.00±0.00	80.4±1.7
`doubao-seed-2.0-pro`	0.5±2.0	6.2±3.1	11.4±3.5	13.2±5.7	7.8±1.9	14.6±1.7	0.01±0.01	0.00±0.00	87.2±1.2
`deepseek-v4-pro` (thinking off)	25.0±3.6	25.9±1.7	44.1±1.9	35.4±2.6	32.6±0.5	33.4±3.6	0.01±0.01	0.25±0.11	64.9±4.7
`qwen3.6-plus` (thinking off)	17.6±11.1	24.2±7.0	68.7±1.9	23.9±3.1	33.6±3.3	44.2±1.9	0.18±0.10	0.08±0.01	56.0±3.6
`qwen3.7-plus` (thinking off)	12.4±12.3	23.1±8.4	67.5±0.8	31.6±1.5	33.7±4.6	45.6±1.7	0.17±0.11	0.06±0.01	55.3±4.1
`qwen3.8-max` (thinking off)	42.6±4.1	45.8±4.3	70.2±5.6	47.8±5.0	51.6±2.4	58.7±1.5	0.04±0.02	0.03±0.01	42.0±1.0
`Qwen3-1.7B`	4.7±1.5	8.3±2.5	32.0±2.1	2.3±5.0	11.8±1.5	11.3±0.3	0.52±0.07	0.29±0.01	78.6±0.6
`Qwen3-4B`	6.1±0.3	25.8±1.7	38.3±0.5	10.3±1.7	20.1±0.3	18.1±0.0	0.06±0.02	0.24±0.02	83.6±0.6
`Qwen3-8B`	10.1±0.9	27.9±1.1	45.5±1.3	17.1±2.4	25.2±0.7	28.4±0.9	0.30±0.03	0.18±0.02	67.2±0.6
`Qwen3-14B`	14.7±1.8	25.9±3.5	52.4±1.1	24.1±4.2	29.3±0.5	41.9±0.7	0.18±0.04	0.17±0.04	56.4±0.6
`Qwen3-30B-A3B`	16.9±2.2	21.6±1.0	40.2±0.3	13.3±2.4	23.0±0.6	21.3±0.3	0.19±0.00	0.16±0.01	76.5±0.5
Task-Trained LLM Policies (`Qwen3-8B`)									
SFT	52.5±1.2	63.9±0.4	59.7±1.5	5.7±1.4	45.5±0.8	72.7±0.2	0.02±0.00	0.10±0.03	58.7±1.3
GRPO	17.8±2.2	33.6±1.2	47.7±1.3	17.6±5.9	29.2±1.4	33.3±0.7	0.19±0.03	0.17±0.02	64.4±0.7
SFT → GRPO	59.2±1.9	63.8±0.3	72.5±2.6	28.6±1.8	56.0±0.7	74.5±0.3	0.08±0.01	0.12±0.01	23.4±0.3

hyperparameters in Appendix B.8. We also report two privileged references: HYDRA-Teacher, the planner from Section 3.2 supplied with the true observation parameters, and the executable clairvoyant policy that can read the hidden patient state.

Evaluation. We run experiments across three random seeds. Our primary metric, SonoScore, normalizes the terminal return in Eq. (1) between an abstaining baseline (0) and a case-specific clairvoyant ceiling (100), which assumes first-look success from every selected adequate window calculated from the ground truth report. We also report evidence coverage (*Cov*), the fraction of diagnostic questions supported by a usable plane from an adequate window in the trajectory. For structured findings, we report the numbers of unsupported negative and unsupported positive verdicts (*U.N./U.P.*) per examination, together with the abstention rate (*Abst.*) defined as the percentage of findings reported as indeterminate. For free-text reports, we measure unsupported negatives (*U.N.*), contradictions (*Contr.*), and undeclared claims (*U.C.*) compared with the scanning trajectory and structured findings. These metrics capture distinct forms of hallucination in the generated report. We also monitor documentation quality (*Q.C.*) using the official clinical quality-control checklist (Appendix C). Full metric definitions and computation details are in Appendix B.7.

5 RESULTS

Results show that adaptation matters when observations redirect acquisition, whereas re-acquisition suffices when the relevant windows are already covered. Acquisition, structured verdicts, and report faithfulness do not always improve together. Table 1 and Figure 3 summarize these results; examples and supporting analyses are in Appendices F and D.

Adaptation matters when observations change where to look. Across the 26 non-privileged policies, SonoScore correlates strongly with evidence coverage ($r = 0.90$; Figure 3a). The fixed protocol scores 25.8 and produces 1.15 unsupported negative verdicts per examination. Re-looking at unusable views and abstaining without evidence raises SonoScore to 46.5 and eliminates these er-

Table 2: **Report generation on SONOARENA.** Green (1st), purple (2nd), and yellow (3rd) backgrounds are to distinguish the top three results.

Policies	U.N. by exam. family ↓ (based on report)				Overall (based on report)			
	Thyroid	Breast	Liver	Focused em.	U.N.↓	Contr. ↓	U. C. ↓	Q.C. ↑
General-Purpose LLM Policies								
`gpt-4o-mini`	0.04	0.10	0.01	0.15	0.07	1.8	0.6	66.5
`gpt-5.4-mini`	1.83	0.79	0.69	0.63	1.04	15.4	9.1	70.9
`gpt-5.5`	0.66	0.58	0.34	1.58	0.74	26.5	5.1	71.1
`claude-sonnet-4-6`	1.03	0.75	0.51	1.38	0.87	20.7	5.1	75.0
`claude-sonnet-5`	1.72	1.61	1.00	3.30	1.75	34.6	3.5	76.8
`claude-opus-5`	2.52	1.42	1.19	2.05	1.77	33.1	2.4	75.8
`gemini-3.1-pro`	0.19	0.18	0.24	1.57	0.43	10.7	1.9	75.8
`gemini-3.8-flash`	0.34	0.29	0.58	1.12	0.52	15.8	2.8	71.7
`grok-4.6`	0.52	0.28	1.32	3.28	1.35	25.5	1.4	76.7
`glm-5.3`	0.80	0.32	1.20	0.88	0.80	25.3	3.6	77.7
`kimi-k3`	0.15	0.18	0.12	0.73	0.25	19.8	2.6	40.4
`doubao-seed-2.0-pro`	0.12	0.27	0.18	0.80	0.29	20.7	1.9	40.0
`deepseek-v4-pro` (thinking off)	0.73	0.40	0.07	1.80	0.63	8.7	5.3	70.9
`qwen3.6-plus` (thinking off)	0.88	1.35	0.62	2.67	1.43	21.5	4.7	77.8
`qwen3.7-plus` (thinking off)	0.92	1.09	0.47	2.15	1.20	19.5	5.0	77.7
`qwen3.8-max` (thinking off)	0.53	0.50	0.35	1.98	0.68	13.7	1.8	80.0
`Qwen3-1.7B`	0.00	0.01	0.00	0.00	0.00	0.0	0.0	42.5
`Qwen3-4B`	0.96	0.47	0.07	1.16	0.61	16.6	6.6	55.0
`Qwen3-8B`	2.23	0.90	0.11	1.25	1.11	32.0	2.5	62.9
`Qwen3-14B`	0.53	1.00	0.07	1.19	0.64	20.4	1.9	66.0
`Qwen3-30B-A3B`	1.67	1.96	0.25	0.94	1.23	36.8	12.6	66.0
Task-Trained LLM Policies (`Qwen3-8B`)								
SFT	0.00	0.01	0.12	0.01	0.04	36.1	0.8	54.8
GRPO	1.48	0.73	0.10	1.72	0.93	30.7	3.2	62.3
SFT → GRPO	0.01	0.01	0.11	0.04	0.04	3.0	0.8	53.8

rors. This rule instantiates Proposition 3's repeated-acquisition mechanism without deciding which window to select next. Compared with the SFT→GRPO task-trained policy, the protocol performance is close on thyroid (53.6 versus 59.2) and breast (62.4 versus 63.8), but falls behind on liver (65.2 versus 72.5) and especially focused emergency (4.8 versus 28.6), where closed or competing windows require acquisition to redirect. This pattern agrees with Propositions 1–2 that local re-acquisition suffices in low-divergence families, while high-divergence families require observation-conditioned window selection.

Task-trained policies first acquire evidential discipline. Zero-shot `Qwen3-8B` scores 25.2, covers only 28.4% of questions, and produces 0.30 unsupported negative verdicts per examination. SFT raises SonoScore to 45.5 and coverage to 72.7%, while reducing this error rate to 0.02. Applying GRPO after SFT further raises SonoScore to 56.0, above the re-look rule (46.5) but below the teacher (66.8). In contrast, applying GRPO directly to the base model yields only a modest improvement to 29.2. The benefit of verifier feedback therefore depends on first learning re-acquisition and evidential licensing through imitation. The remaining gap is largest on focused emergency (28.6 versus 53.9), suggesting that belief-conditioned window selection is not resolved by either training stage. Further ablations are provided in Appendix D.1.

General-purpose LLM-based policies show the complementary failure. Overall SonoScore among general-purpose LLM-based policies ranges from 7.8 for `doubao-seed-2.0-pro` to 55.1 for `gpt-5.5`. The best policies surpass the re-look rule overall, but remain below the task-trained policy and teacher. On focused emergency, however, `gemini-3.8-flash` and `claude-opus-5` score 54.6 and 64.2, exceeding the teacher's 53.9, while lagging the re-look rule on protocol-complete families. These agents can conduct highly adaptive examinations, including redirecting acquisition in response to observations, but are less reliable at executing a regular scanning protocol.

Structured findings and reports follow different trends. Improvements in structured findings do not transfer uniformly to the generated report. From zero-shot `Qwen3-8B` to SFT, SonoScore rises from 25.2 to 45.5, coverage from 28.4% to 72.7%, and structured unsupported negatives fall from 0.30 to 0.02. Over the same transition, documentation quality falls from 62.9 to 54.8. Adding GRPO further raises SonoScore to 56.0, while documentation quality remains at 53.8. Structured findings and free-text reports therefore require separate evaluation.

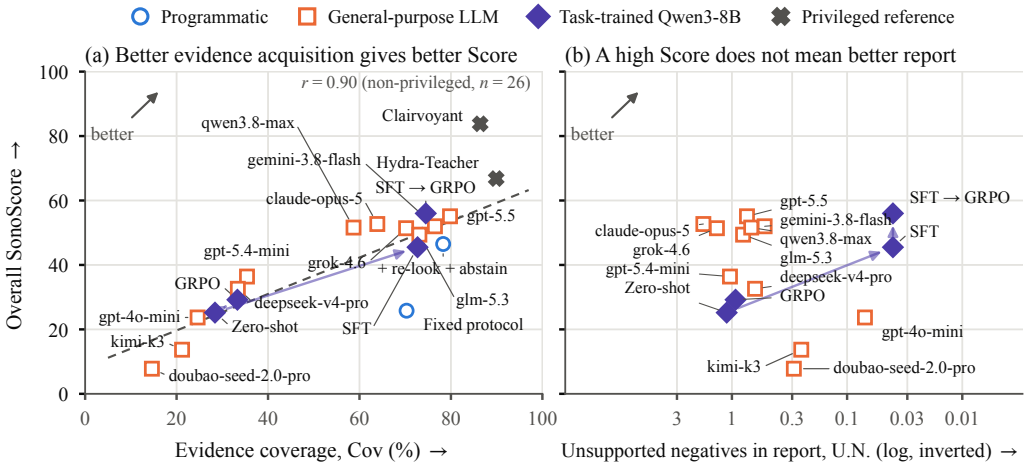


Figure 3: **SonoScore reflects acquisition, but not necessarily report faithfulness. (a)** SonoScore against evidence coverage, with a least-squares fit over non-privileged policies. **(b)** SonoScore against unsupported negatives in free-text reports (U.N.; log scale, inverted). Arrows trace two training paths from zero-shot `Qwen3-8B`: directly to GRPO, and through SFT to SFT→GRPO.

Frontier LLM agents can hallucinate despite faithful findings. Several frontier agents acquire strong evidence and produce well-supported structured findings, but overclaim when writing the report. `gpt-5.5`, for example, achieves 79.8% coverage and produces no unsupported structured verdicts, yet its report contains 0.74 unsupported negatives per examination, contradicts 26.5% of declined questions, and introduces 5.1% undeclared claims. Policies based on smaller models show the opposite pattern. `gpt-4o-mini` has only 0.07 report-level unsupported negatives and `Qwen3-1.7B` records none, but their coverage is only 24.6% and 11.3%, respectively; both abstain on most questions. Their lower hallucination rates therefore reflect fewer claims, rather than more complete or faithful reports.

Adaptive acquisition remains effective under test-time shifts. To test whether the benefit of adaptation depends on the nominal simulator parameters, we rerun focused emergency after perturbing the miss rate and usable-plane probability. Adaptive policies retain their advantage across these shifts, whereas the fixed protocol is more sensitive to changes in window usability. This supports the regime distinction in Propositions 1–2. Full results are provided in Appendix D.

Table 3: SonoScore of focused emergency under test-time shifts (seed 1).

Policy	In-dist.	Reveal ×0.5	Reveal ×1.5	Plane 0.7	Plane 0.5
Fixed protocol	4.0	1.4	10.4	−13.8	−24.4
Re-look + abstain	4.3	0.3	7.3	4.1	7.7
HYDRA-Teacher	54.8	34.1	62.4	55.7	49.4
SFT → GRPO	29.0	27.1	36.2	28.3	22.8

6 CLINICIAN REVIEW AND FUTURE WORK

A clinician rated 40 completed episodes across the 4 families, and the ratings generally agreed with the scorer ($\rho = -0.52$). This provides an initial audit of clinical utility (Appendix D.3). The remaining gap between trained and reference policies suggests room for further improvements in agentic examination planning. Recent progress in ultrasound image perception and robotic probe control complements this work. Combining these components with adaptive acquisition and reporting could support an automated, closed-loop agentic ultrasound system.

7 CONCLUSION

Ultrasound examination is an agentic, closed-loop task: an agent must acquire evidence, judge when it is sufficient, and report only licensed claims. SONOARENA turns this process into an interactive benchmark with separate evaluation of acquisition, structured findings, and report faithfulness. Our analysis and experiments identify when fixed protocols suffice and when observation-conditioned redirection requires adaptation. This makes adaptive acquisition a measurable target for future agentic ultrasound systems.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, LaTeX formatting, and assistance in drafting prose. We also used Claude to produce the sampling specification for the synthetic focused-emergency family. During evaluation, frozen LLMs (`gpt-4o-mini` by default and `claude-sonnet-4-6` for reports) assigned report clauses to examination windows and polarities before deterministic rule-based scoring. We reviewed all AI-assisted text and research artifacts and take responsibility for the final content, claims, methods, and artifacts.

ETHICS STATEMENT

SONOARENA uses a public structured ultrasound-report dataset for three report-grounded families and a synthetically generated focused-emergency family; it introduces no private dataset. Released benchmark cases contain derived symbolic fields rather than source report text or identifiers, and players do not receive patient images or source reports. Clinical indications and acoustic-access obstacles are generated, and the focused-emergency family uses a Claude-produced sampling specification. Nothing here is a clinical validation or a deployable device, and we do not claim diagnostic performance on real ultrasound studies.

REPRODUCIBILITY STATEMENT

The benchmark's POMDP formulation, symbolic observation model, reward, interaction protocol, and reference baselines are specified in § 2 and § 3.2. Case-generation parameters, the utility matrix, prompts, and decoding settings are to be released in the appendix and in anonymized supplementary code. Because (y, Ω) is known to the sandbox, all headline metrics (regret, unsupported-negative rate, premature termination) are deterministic functions of the report and the ground-truth state, with no LLM-as-judge.

REFERENCES

Seth Kofi Abrokwa, Lisa C. Ruby, Charlotte C. Heuvelings, and Sabine Bélard. Task shifting for point of care ultrasound in primary healthcare in low- and middle-income countries—a systematic review. *eClinicalMedicine*, 45:101333, 2022. doi: 10.1016/j.eclinm.2022.101333. URL `https://doi.org/10.1016/j.eclinm.2022.101333`.

Camilla Aakjær Andersen, Sinead Holden, Jonathan Vela, Michael Skovdal Rathleff, and Martin Bach Jensen. Point-of-care ultrasound in general practice: A systematic review. *The Annals of Family Medicine*, 17(1):61–69, 2019. doi: 10.1370/afm.2330. URL `https://doi.org/10.1370/afm.2330`.

Yunke Ao, Masoud Moghani, Mayank Mittal, Manish Prajapat, Luohong Wu, Frederic Giraud, Fabio Carrillo, Andreas Krause, and Philipp Fürnstahl. SonoGym: High performance simulation for challenging surgical tasks with robotic ultrasound. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Khadija Azhar, Byoung-Dai Lee, Shi Sub Byon, SeungJae Lee, Kyu Ran Cho, and Sung Eun Song. Semiautomated breast ultrasound report generation using multimodal large language models and deep learning. *Frontiers in Medicine*, 13:1679203, 2026.

Yuan Bi, Yiping Zhou, Pei Liu, Feng Li, Zhongliang Jiang, and Nassir Navab. From scanning guidelines to action: A robotic ultrasound agent with LLM-based reasoning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume LNCS 16893. Springer Nature Switzerland, September 2026. URL `https://papers.miccai.org/miccai-2026/0406-Paper2814.html`.

Yicheng Gao, Xiaolin Zhou, Yahan Li, Yue Zhao, and Ruishan Liu. Medexagent: Training llm agents to ask, examine, and diagnose in noisy clinical environments. *arXiv preprint arXiv:2605.07058*, 2026.

Peixuan Ge, Tongkun Su, Faqin Lv, Baoliang Zhao, Peng Zhang, Chi Hong Wong, Liang Yao, Yu Sun, Zenan Wang, Pak Kin Wong, et al. Ultrasound report generation with multimodal large language models for standardized texts. *arXiv preprint arXiv:2505.08838*, 2025.

Haojun Jiang, Andrew Zhao, Qian Yang, Xiangjie Yan, Teng Wang, Yulin Wang, Ning Jia, Jiangshan Wang, Guokun Wu, Yang Yue, et al. Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system. *Nature Communications*, 16(1):7893, 2025.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Anjie Le, Henan Liu, Yue Wang, Zhenyu Liu, Rongkun Zhu, Taohan Weng, Jinze Yu, Boyang Wang, Yalun Wu, Kaiwen Yan, Quanlin Sun, Meirui Jiang, Jialun Pei, Siya Liu, Haoyun Zheng, Zhoujun Li, J. Alison Noble, Jacques Souquet, Xiaoqing Guo, Manxi Lin, and Hongcheng Guo. U2-BENCH: Benchmarking large vision–language models on ultrasound understanding. In *International Conference on Learning Representations*, 2026.

Jun Li, Tongkun Su, Baoliang Zhao, Faqin Lv, Qiong Wang, Nassir Navab, Ying Hu, and Zhongliang Jiang. Ultrasound report generation with cross-modality feature alignment via unsupervised guidance. *IEEE Transactions on Medical Imaging*, 44(1):19–30, 2025. doi: 10.1109/TMI.2024.3424978.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.

Zichen Liu, Anya Sims, Keyu Duan, Changyu Chen, Simon Yu, Xiangxin Zhou, Haotian Xu, Shaopan Xiong, Bo Liu, Chenmien Tan, Weixun Wang, Hao Zhu, Weiyan Shi, Diyi Yang, Michael Qizhe Shieh, Yee Whye Teh, Wee Sun Lee, and Min Lin. GEM: A gym for generalist LLMs. In *ICLR*, 2026.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, et al. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

Giuseppe A Orlando, Paolo Papotti, Maria A Zuluaga, Olivier Humbert, and Marco Lorenzi. Medscribe: Clinically grounded ct reporting through agentic workflows. *arXiv preprint arXiv:2605.01779*, 2026.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wołczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. BALROG: Benchmarking agentic LLM and VLM reasoning on games. In *ICLR*, 2025.

Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. Training software engineering agents and verifiers with SWE-Gym. In *ICML*, 2025.

Pengcheng Qiu, Chaoyi Wu, Junwei Liu, Qiaoyu Zheng, Yusheng Liao, Haowen Wang, Yun Yue, Qianrui Fan, Shuai Zhen, Jian Wang, Jinjie Gu, Yanfeng Wang, Ya Zhang, and Weidi Xie. Evolving interactive diagnostic agents in a virtual clinical environment. *arXiv preprint arXiv:2510.24654*, 2025.

Roy Radner and Joseph E. Stiglitz. A nonconcavity in the value of information. In Marcel Boyer and Richard E. Kihlstrom (eds.), *Bayesian Models in Economic Theory*, pp. 33–52. North-Holland, 1984.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal benchmark for tool-using clinical ai agents. *npj Digital Medicine*, 9:499, 2026. doi: 10.1038/s41746-026-02674-7. URL https://doi.org/10.1038/s41746-026-02674-7.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *ICLR*, 2021.

Ruud J. G. van Sloun. Active inference and deep generative modeling for cognitive ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 71(11):1478–1490, 2024. doi: 10.1109/TUFFC.2024.3466290.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *EMNLP*, 2022.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN: Understanding self-evolution in LLM agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025.

Taohan Weng, Chi Zhang, Chaoran Yan, Siya Liu, Xiaoyang Liu, Yalun Wu, Boyang Wang, Boyan Wang, Jiren Ren, Kaiwen Yan, Jinze Yu, Kaibing Hu, Henan Liu, Haoyun Zheng, Anjie Le, and Hongcheng Guo. Dolphin v1.0 technical report. *arXiv preprint arXiv:2509.25748*, 2025. URL https://arxiv.org/abs/2509.25748.

Juncheng Wu, Letian Zhang, Yuhan Wang, Haoqin Tu, Hardy Chen, Zijun Wang, Cihang Xie, and Yuyin Zhou. Clinseekagent: Automating multimodal evidence seeking for agentic clinical reasoning. *arXiv preprint arXiv:2605.20176*, 2026.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Jiaqi Liu, Honglin Guo, Yajie Yang, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, Tao Gui, Zuxuan Wu, Qi Zhang, Xuanjing Huang, and Yu-Gang Jiang. AgentGym-RL: An open-source framework to train LLM agents for long-horizon decision making via multi-turn RL. In *International Conference on Learning Representations*, 2026. Oral presentation.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS*, 2022.

Yang Yue, Yulin Wang, Haojun Jiang, Pan Liu, Shiji Song, and Gao Huang. EchoWorld: Learning motion-aware world models for echocardiography probe guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25993–26003, 2025.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *ICLR*, 2024.

Appendix Contents

A Related work **14**

B Formulation and implementation details **15**

C Ultrasound examination quality rubric **30**

D Additional results **32**

E Sensitivity to utility and simulator assumptions **35**

F Example trajectories and reports **36**

A Related work

Ultrasound interpretation and acquisition. Ultrasound AI has developed along two main directions. Vision and vision–language models interpret pre-acquired images or videos for tasks such as classification, segmentation, and report generation (Le et al., 2026; Weng et al., 2025; Ge et al., 2025; Azhar et al., 2026). A separate line of work assists image acquisition through probe guidance, robotic control, and standard-plane navigation (Yue et al., 2025; Ao et al., 2025; Jiang et al., 2025). Recent systems use language models to translate scanning guidelines into adaptive robotic actions (Bi et al., 2026), while cognitive ultrasound frames low-level acquisition as active information seeking (van Sloun, 2024). These approaches address individual parts of an ultrasound examination. They do not evaluate the full sequence from diagnostic evidence acquisition through stopping and reporting.

Interactive medical agents. Interactive medical agents gather evidence over multiple turns. Existing systems ask patients follow-up questions (Li et al., 2024), request diagnostic tests (Nori et al., 2025; Gao et al., 2026), navigate clinical records and medical tools (Wu et al., 2026), or coordinate simulated clinical encounters (Schmidgall et al., 2026). Agentic reporting systems also gather localized evidence from an already acquired imaging volume before generating a report (Orlando et al., 2026). DiagGym simulates examination results from EHRs, and DiagAgent uses multi-turn reinforcement learning to decide which examination to order next and when to finalize a diagnosis (Qiu et al., 2025). Its actions select high-level clinical tests and receive textual results; it does not model acquisition decisions within an imaging modality, patient-specific acoustic access, or whether an observation licenses a report claim. These agents either interact with information systems at the clinical-workflow level or reason over images acquired by someone else. SonoArena targets decisions within the imaging examination and models the acquisition process.

Interactive environments for agent evaluation and learning. Interactive environments evaluate language models as policies whose actions affect subsequent observations. Early benchmarks established this setting across embodied, scientific, and web tasks (Shridhar et al., 2021; Wang et al., 2022; Yao et al., 2022; Zhou et al., 2024). Recent work includes long-horizon game environments such as BALROG (Paglieri et al., 2025), executable repository-level software engineering in SWE-Gym (Pan et al., 2025), and multi-turn policy learning through AgentGym-RL, RAGEN, and GEM (Xi et al., 2026; Wang et al., 2025; Liu et al., 2026). SonoArena extends this paradigm to a medical sensing task whose actions represent physical acquisition decisions. While abstracting away low-level perception and probe control, it combines adaptive view and measurement selection, stopping, and evidence-grounded reporting as a single policy.

B FORMULATION AND IMPLEMENTATION DETAILS

This appendix specifies the observation model and reward used in § 2, proves Propositions 1–3, and documents the benchmark implementation. We state explicitly when an implementation choice is narrower than the general formulation.

B.1 THE OBSERVATION MODEL

Information state. In a POMDP (Kaelbling et al., 1998), the agent chooses an action a and observes o with probability $\Pr(o \mid z, x, a)$ under latent patient variables z and observable examination state x. The history updates the information state $I_t = (b_t(z), x_t)$, which is sufficient for choosing the next action. An action can therefore be useful because of the information it reveals, even before it changes the terminal reward. In ultrasound, the result from one window can change both the next window and the decision to stop.

Why the observation model sets the false-positive rate to zero. We set $\alpha_i = 0$ as a simplifying assumption. Ultrasound can produce false-positive signs: side-lobe and slice-thickness artifacts can mimic sludge, while mirror-image or reverberation artifacts can mimic a collection or mass. The report-grounded corpora record final findings, not artifact events, so they cannot support an estimate of a per-finding α_i. The analysis instead concerns whether an adequate view supports a negative and whether another acquisition repays its cost, both of which depend on β_i and Ω. With $\alpha_i = 0$, one observed sign is decisive and the environment cannot reward a confirmatory view of a positive. Behavior after a positive is therefore not clinically calibrated. Misses ($\beta_i > 0$) remain retryable: a finding missed on one pass may appear on the next. This separates a stochastic miss from persistent inaccessibility encoded by ω.

Work-up actions. DETAIL, MEASURE, and CAPTURE cost probe time and unlock morphology, dimensions, and stored planes for the report. They affect the separate documentation monitor (QC), not the structured diagnostic utility u; this keeps report documentation distinct from the question of whether the trajectory licensed each diagnostic verdict.

The likelihood ratio at a closed window. From the third row of the observation model,

$$\Lambda_{\omega_w=0} \;=\; \frac{\Pr(\text{no sign} \mid y_i{=}1, \omega_w{=}0)}{\Pr(\text{no sign} \mid y_i{=}0, \omega_w{=}0)} \;=\; 1,$$

exactly, whereas at an adequate window $\Lambda_{\omega_w=1} = \beta_i$, materially below one for any useful test. An agent that does not carry Ω applies the second ratio in both cases. The defect is in the state space, not in the parameters, and the source reports do not identify which mechanism produced a negative finding.

B.2 REWARD: BOUNDARIES, ORDERING AND MAGNITUDES

Boundaries of the definition. An ABSENT verdict from an adequate window that missed the finding is licensed but incorrect, and is scored as a false negative. Calling it unsupported would build a required number of looks into the checker, although choosing how many times to look is part of the policy. A PRESENT verdict without support from a scanned window in $\Omega_{\cdot i}$ is scored as a false positive even when it matches the hidden state. The action cost $c(a)$ includes both scans and work-up actions in trajectory τ.

The ordering. Writing u_{FN} for a licensed ABSENT on a finding that was present, u_{FP} for a false PRESENT, u^{+}_{IND} and u^{-}_{IND} for INDETERMINATE in the two states, and $\phi < 0$ for the extra charge on an unsupported ABSENT:

$$\underbrace{u_{\text{FN}} < u_{\text{FP}}}_{\text{a miss costs more than a false alarm}}, \qquad \underbrace{u_{\text{FN}} < u^{+}_{\text{IND}} < u_{\text{TP}}}_{\text{honesty beats error, loses to correctness}}, \qquad \underbrace{u_{\text{FP}} < u^{-}_{\text{IND}} < u_{\text{TN}}}_{\text{likewise on the negative side}},$$

$$\phi \;\le\; \min\left(u^{+}_{\text{IND}} - u_{\text{FN}},\; u^{-}_{\text{IND}} - u_{\text{TN}}\right).$$

The middle pair is what stops a policy from scoring well by abstaining everywhere. The last line is derived rather than chosen: require that an unsupported ABSENT never have higher expected utility than INDETERMINATE at any prior $p \in [0,1]$ on y_i, and compare the two expected utilities at the endpoints $p = 1$ (where the unsupported negative earns $u_{\text{FN}} + \phi$ against u^{+}_{IND}) and $p = 0$ (where it earns $u_{\text{TN}} + \phi$ against u^{-}_{IND}); both expected utilities are linear in p, so the two endpoint constraints suffice.

Default magnitudes. We express all utilities in units of one unnecessary biopsy by setting $u_{\text{FP}} = -1$. The default false-negative magnitude corresponds to an illustrative biopsy threshold $p^\star = 0.05$: indifference at that threshold gives $|u_{\text{FN}}|/|u_{\text{FP}}| = (1 - p^\star)/p^\star = 19$. This threshold is a reward-design choice rather than an estimate of clinical prevalence or a family-specific recommendation. The claims rely on the utility ordering above; Appendix E varies the magnitudes and the observation parameters directly.

B.3 Proposition 1: statement and proof

From assumptions to an open-loop trajectory. Proposition 1 uses the two assumptions stated in the main text: a narrow prior and a reliable observation channel. Let δ be the prior mass outside the modal state, and let β bound the probability that a look departs from its noiseless outcome. We show that the optimal policy follows one fixed action sequence and stops at one fixed step on an event of probability at least $1 - \delta - N\beta$, where N counts finding-specific observations along its noiseless trajectory. Observation-independent actions and termination are consequences on this event, not additional hypotheses. For later reference, we label no branch points as (C1), the narrow prior as (C2), the reliable channel as (C3), and fixed termination as (C4). Only (C2) and (C3) are assumptions of the proposition.

Setting and notation. We recall the objects of § 2.1 and § 2.1 and fix the notation the proof uses.

- *State.* $s = (y, \Omega)$, where $y \in \{0,1\}^m$ records which of the m findings are present and $\Omega \in \{0,1\}^{|\mathcal{W}| \times m}$ records which windows are adequate for which findings. The state is drawn once from the prior b_0 and does not change during the examination. The set of states is finite.
- *Actions and observations.* An action a is a scan of a window, a work-up on a scanned window, or REPORT. An observation o is what the sandbox returns for that action: for a scan, a sign or no sign for each finding the window's anatomy covers, plus the adequacy cue.
- *History.* $h_t = (a_1, o_1, \ldots, a_t, o_t)$ is the record of the first t actions and their observations; h_0 is the empty record.
- *Policy.* A policy π is a rule that reads the history h_{t-1} and returns the next action a_t. The episode ends at the first REPORT; we call that step the *stopping time* T^π and require $T^\pi \le H$, where H is the horizon, the probe-time budget counted in actions. The report $\hat{r}$ is written from the final history.
- *Return.* $R = \sum_{i=1}^{m} u(\hat{r}_i, y_i, \Omega_{\cdot i}, h_T) - \lambda \sum_{a \in \tau} c(a)$, the reward of § 2.1: u scores the verdict $\hat{r}_i$ on finding i against the truth y_i, the adequate windows $\Omega_{\cdot i}$ and the evidence recorded in h_T; $c(a)$ is the probe time of action a, summed over the trajectory τ, and λ is its weight. R depends on the state, the interaction history, and the report.
- *Span.* Δ is the largest possible difference between two returns: the range of $\sum_i u$ plus λ times the largest total cost a trajectory of at most H actions can incur. It is finite because u takes finitely many values and costs are bounded.
- *Value.* $V^\pi = \mathbb{E}[R \mid \pi, b_0]$ is the expected return of π when the state is drawn from b_0 and the observations from the channel of § 2.1. A policy $\pi^\star$ is *optimal* if $V^{\pi^\star} \ge V^\pi$ for every π; we write $V^\star = V^{\pi^\star}$. An optimal policy exists because the state, action and observation sets are finite and the horizon is bounded.

An *open-loop* policy ignores observations during acquisition. It executes a fixed action sequence $\bar{a}_1, \ldots, \bar{a}_T$ and stops at a fixed step T, whatever it sees; only its report may depend on the history. This matches the fixed-protocol baseline of § 4, which scans the station list in order and then reports

its observations. The report remains history-dependent because only acquisition and stopping are open loop.

Quantified conditions.

- (C2$_\delta$) There is a state $s^\star = (y^\star, \Omega^\star)$, the *modal state*, with $b_0(s^\star) \geq 1 - \delta$. So the prior puts at most δ of its mass on states other than $s^\star$.
- (C3$_\beta$) Fix the state to be $s^\star$. Consider a scan of window w and a finding i that w is adequate for ($\Omega^\star_{wi} = 1$). Call the observation on i *noiseless* if it is a sign when $y^\star_i = 1$ and no sign when $y^\star_i = 0$, and say the look *departs* otherwise. We require that each such look departs with probability at most β, and that departures are the only randomness in the channel. In the observation model of § 2.1 a look can depart in two ways: a present finding is missed, which happens with probability β_i, or the window returns no usable plane through a transient failure, which happens with probability p_f; so $\beta = \max_i \beta_i + p_f$ suffices. Two things are deterministic under this model and therefore never depart: an absent finding never shows a sign, because the false-positive rate α is zero, and a closed window ($\Omega^\star_{wi} = 0$) always returns "no usable plane." We do *not* assume that departures are independent across the findings seen in one look; a plane failure blanks all of them at once.

The noiseless trajectory and the open-loop competitor. Fix an optimal policy $\pi^\star$. Imagine running it in the modal state $s^\star$ with every look returning its noiseless outcome. Then each action has exactly one possible observation, so the run has exactly one possible history: starting from h_0, $\pi^\star$ picks $\bar{a}_1$, the noiseless observation $\bar{o}_1$ follows, $\pi^\star$ reads $(\bar{a}_1, \bar{o}_1)$ and picks $\bar{a}_2$, and so on until it picks REPORT. Call this history the *noiseless trajectory*

$$\bar{h} = (\bar{a}_1, \bar{o}_1, \ldots, \bar{a}_{\bar{T}}, \bar{o}_{\bar{T}}), \qquad \bar{a}_{\bar{T}} = \text{REPORT},$$

and write $\bar{h}_t$ for its first t steps. A *finding-look* along $\bar{h}$ is a pair (step t, finding i) such that $\bar{a}_t$ scans a window w with $\Omega^\star_{wi} = 1$; these are exactly the places where a departure is possible. Let N be their number; since each step looks at no more than m findings, $N \leq \bar{T}m \leq Hm$.

Define the open-loop policy π_{OL} as follows: execute $\bar{a}_1, \ldots, \bar{a}_{\bar{T}}$ in order regardless of what is observed, and at step $\bar{T}$ write the report that $\pi^\star$ would write when handed the history π_{OL} actually accumulated. This is a legitimate open-loop policy: its action sequence and its stopping time are fixed in advance, and only the report reads the history.

Proposition 1. *Under* (C2$_\delta$) *and* (C3$_\beta$)*,*

$$V^\star - V^{\pi_{\mathrm{OL}}} \;\leq\; (\delta + N\beta)\,\Delta.$$

In particular, when $\delta = \beta = 0$ the open-loop policy π_{OL} is optimal. Moreover, there is an event E with $\Pr(E) \geq 1 - \delta - N\beta$ on which the optimal policy $\pi^\star$ takes a fixed action at every step (C1) *and stops at a fixed step* (C4)*.*

Proof. The argument has four steps: put both policies on one probability space; name the event E on which nothing goes wrong and bound its complement; show that on E the two policies do exactly the same thing and so earn the same return; and bound the loss off E by Δ.

Step 1: one probability space for both policies. The randomness of an episode has two sources: the state s, drawn from b_0, and the observation noise, which by (C3$_\beta$) consists only of departures. Draw $s \sim b_0$ and, for each step $t \leq H$, an independent random variable ξ_t uniform on $[0, 1]$. Define the observation at step t to be a fixed function of three things: the state s, the action taken at step t, and ξ_t. The function is chosen so that for every scan action and every finding i the window is adequate for, the set of ξ_t-values that make the look on i depart has probability equal to the channel's departure probability, which under $s^\star$ is at most β by (C3$_\beta$). Such a function exists because any finite-valued distribution can be sampled from a uniform variable. Write $D_{t,i}$ for the event "the look on finding i at step t departs." The point of this construction is that the law of $(s, \xi_1, \ldots, \xi_H)$ does not depend on which policy is run, so we may run $\pi^\star$ and π_{OL} on the same draw and compare their returns pathwise.

Step 2: the good event and its probability. Let

$$E \;=\; \{s = s^\star\} \;\cap \bigcap_{(t,i)} D^c_{t,i},$$

where the intersection runs over the N finding-looks of the noiseless trajectory $\bar{h}$. In words, E is the event that the state is the modal one and none of the looks that $\bar{h}$ makes departs. Its complement is contained in the union of $\{s \neq s^\star\}$ and the N events $D_{t,i}$, so

$$\Pr(E^c) \;\le\; \Pr(s \neq s^\star) + \sum_{(t,i)} \Pr(D_{t,i} \cap \{s = s^\star\}) \;\le\; \delta + N\beta,$$

where the first term is at most δ by (C2$_\delta$) and each of the N terms is at most β by (C3$_\beta$). This is a union bound; it needs no independence among the $D_{t,i}$.

Step 3: on E, both policies follow $\bar{h}$. We show by induction on t that on E the history after t steps is $\bar{h}_t$ for both policies. For $t = 0$ both histories are empty, which is $\bar{h}_0$. Suppose the claim holds after $t-1$ steps. The optimal policy $\pi^\star$ reads $\bar{h}_{t-1}$ and returns $\bar{a}_t$, because $\bar{a}_t$ was defined as the action $\pi^\star$ takes on $\bar{h}_{t-1}$. The open-loop policy π_{OL} also takes $\bar{a}_t$, because that is its t-th scheduled action. So both take the same action. On E the state is $s^\star$ and no look along $\bar{h}$ departs, so the observation returned for $\bar{a}_t$ is its noiseless outcome in $s^\star$, which is $\bar{o}_t$ by definition of $\bar{h}$; the two policies receive the same observation because they took the same action on the same draw. Hence both histories become $\bar{h}_t$, completing the induction. In particular both policies reach $\bar{h}_{\bar{T}}$, and both stop at step $\bar{T}$: $\pi^\star$ because $\bar{a}_{\bar{T}} = \textsc{report}$ is what it does on $\bar{h}_{\bar{T}-1}$, and π_{OL} because $\bar{T}$ is its scheduled stopping step.

The returns agree on E: the state is the same, the action sequence is the same so the cost term $\lambda \sum_{a \in \tau} c(a)$ is the same, and the report is the same because π_{OL} was defined to write $\pi^\star$'s report on the history it obtained, which is the history $\pi^\star$ obtained. Thus every term of R agrees and $R^{\pi^\star} = R^{\pi_{\mathrm{OL}}}$ on E. This also establishes the sense in which (C1) and (C4) hold on E: the history of $\pi^\star$ on E is the one fixed sequence $\bar{h}$, so the action it takes at each step and the step at which it stops are constants, not functions of what was observed.

Step 4: off E, and the bound. Off E we know nothing about what the two policies do, but any two returns differ by at most Δ, by definition of Δ. Splitting the expectation over E and E^c and using that the difference vanishes on E by Step 3,

$$\begin{aligned} V^\star - V^{\pi_{\mathrm{OL}}} &= \mathbb{E}\big[(R^{\pi^\star} - R^{\pi_{\mathrm{OL}}})\,\mathbf{1}_E\big] + \mathbb{E}\big[(R^{\pi^\star} - R^{\pi_{\mathrm{OL}}})\,\mathbf{1}_{E^c}\big] \\ &= 0 + \mathbb{E}\big[(R^{\pi^\star} - R^{\pi_{\mathrm{OL}}})\,\mathbf{1}_{E^c}\big] \;\le\; \Delta\,\Pr(E^c) \;\le\; (\delta + N\beta)\,\Delta, \end{aligned}$$

the last inequality being Step 2. When $\delta = \beta = 0$ the right-hand side is zero, so $V^{\pi_{\mathrm{OL}}} \ge V^\star$; since $V^\star$ is the maximum, $V^{\pi_{\mathrm{OL}}} = V^\star$ and π_{OL} is optimal. If one is willing to assume departures independent across finding-looks, the same argument with $\Pr(E) = (1-\delta)(1-\beta)^N$ in Step 2 gives the sharper $\Pr(E^c) \le 1 - (1-\delta)(1-\beta)^N$. □

Remark 1 (what each condition buys). The proof uses (C2$_\delta$) once, to bound $\Pr(s \neq s^\star)$, and (C3$_\beta$) once, to bound the probability that a look in $s^\star$ departs. The adequacy clause of (C3), that the protocol's windows are open, does *not* enter: a window that is closed in $s^\star$ is closed in the state the optimal policy plans for, so $\pi^\star$ never schedules it and the noiseless trajectory never visits it. Closed windows therefore cost an open-loop policy nothing by themselves. They cost it something only when *which* windows are closed differs from case to case, and that is a failure of (C2) on the Ω component of the state, since Ω is part of s. This is the formal reason § 3.1 counts ρ_Ω, the fraction of cases in which a protocol window is closed, as a source of divergence: it measures how often the realized Ω differs from the $\Omega^\star$ the protocol was written for.

Remark 2 (the bound is tight in δ). The linear dependence on δ is not an artifact of the proof; here is an instance in which every open-loop policy loses $\Theta(\delta)$. Take two states. In $s^\star$, which has prior mass $1-\delta$, a finding i is absent. In s', which has prior mass δ, i is present and is adequate only at a window w' that the protocol does not visit; and a second finding j, absent in $s^\star$ and present in s', is adequate at a window w_0 that the protocol does visit. Set $\beta = 0$, so every look is noiseless, and write $g = u_{\mathrm{TP}} - u^{+}_{\mathrm{IND}}$ for the gain from asserting a present finding rather than leaving it indeterminate, and $g^- = u_{\mathrm{TN}} - u^{-}_{\mathrm{IND}}$ for the gain from licensing an absent finding. Assume $g^- < \lambda c(w') < g$, so visiting w' is worthwhile only on the state identified by the cue.

Consider the belief-conditioned policy that scans w_0, visits w' if and only if it saw j there, and then reports. In $s^\star$ it sees no j, skips w', and pays nothing extra. In s' it sees j, which under a noiseless

channel identifies the state as s', visits w', sees i, and asserts i PRESENT from a window adequate for it, earning u_{TP} at cost $\lambda c(w')$.

An open-loop policy fixes its window list in advance, so it either includes w' or it does not. If it always visits w', then in $s^\star$, with probability $1-\delta$, it can earn at most g^- by licensing the negative but pays $\lambda c(w')$, which the belief-conditioned policy did not pay; its expected return is lower by $(1-\delta)(\lambda c(w') - g^-)$. If it never visits w', then in s', with probability δ, it sees j and so knows i is present, but it has no window adequate for i in its scanned set, so a PRESENT would be priced as a false positive by the reward of § 2.1 and its best verdict is INDETERMINATE, worth u^+_{IND}; it loses $g - \lambda c(w')$ relative to the belief-conditioned policy on that state. Its expected return is lower by $\delta\,(g - \lambda c(w'))$. Hence every open-loop policy has regret at least $\min\{(1-\delta)(\lambda c(w') - g^-),\ \delta\,(g - \lambda c(w'))\}$ against the belief-conditioned policy, and therefore at least that against $\pi^\star$. For small δ the minimum is the second term, which is $\Theta(\delta)$.

Remark 3 (what the proposition does not say). It bounds the loss from open-loop *acquisition* against the best policy, not the loss from any particular published station list: π_{OL} is the protocol an oracle who knew $s^\star$ would write, and a real station list may be far from it. It also says nothing about the belief-free closed-loop policy of § 4, which re-looks at empty windows and abstains on thresholds. That policy conditions acquisition on observations, so its comparison with the fixed protocol is not a direct test of Proposition 1. Proposition 2 below removes the first of these limits: it charges every open-loop policy, the oracle's and the published list alike.

B.4 Proposition 2: statement and proof

Why the converse of Proposition 1 is false, and what replaces it. Proposition 1 is a sufficiency result: small δ and β make an open-loop policy near-optimal. The converse fails inside the model of § 2.1. If $\lambda c(w)$ exceeds the largest utility any scan of w can add, the optimal policy reports at once, and that is open-loop for every prior. If every window is closed for every finding on the support of b_0, no policy learns anything and reporting at once is again optimal with δ as large as one likes. And a broad prior whose states all share one noiseless optimal trajectory is served perfectly by that trajectory. So neither (C2) nor (C3) is necessary, and the only condition equivalent to open-loop optimality is (C1)$\wedge$(C4), which is a tautology. What can be proved instead is a *lower bound*: a single branch point, in the sense of (C1), charges every open-loop policy a margin. Remark 2 is one instance; Proposition 2 is the general statement, with hypotheses that name the clinical quantities the margin depends on.

Setting. We work in the setting of Appendix B.3 with a noiseless channel: $\beta = 0$ and $p_f = 0$, so every observation is a fixed function of the state and the action. Remark 5 restores β. Because the state is static and costs are additive, the observation an action returns does not depend on when it is taken, and the return of a history does not depend on the order of its actions except through the report; we use both facts freely. We assume the horizon H leaves room for one action beyond any station list under consideration; the budgets of § 4 do.

Objects and hypotheses. Fix a finding i, a window w', and a window w_0, the *cue window*. Write F' for the set of findings that w' reaches ($A_{w'j} = 1$) and $k' = |F'|$; note $i \in F'$ is allowed. We require:

(H0) *Separation.* w_0 reaches neither i nor any finding in F'. So the utility terms of $F' \cup \{i\}$ do not depend on whether w_0 was scanned, and the observation at w_0 licenses nothing about them.

(H1) *Cue.* The values the noiseless observation at w_0 can take are split into two classes, *branch* and *no-branch*. Define the state events

$$U = \{\text{cue says branch}\} \cap \{y_i = 1,\ \Omega_{w'i} = 1,\ \Omega_{wi} = 0 \text{ for all } w \neq w'\},$$
$$W = \{\text{cue says no-branch}\} \cap \{\Omega_{w'j} = 1 \text{ and } y_j = 0 \text{ for all } j \in F'\},$$

and let $q = \Pr(U)$, $p = \Pr(W)$, $e_U = \Pr(\text{branch}) - q$, $e_W = \Pr(\text{no-branch}) - p$, $e = e_U + e_W$. On U the finding is present and only w' can show it; on W the window is open and shows nothing. The error masses e_U, e_W are the states on which the cue's advice is not backed by the state.

(H2) *Cue window pays for itself.* For every open-loop policy π whose list omits w_0, the open-loop policy π^{+w_0} that prepends $\textsc{scan}(w_0)$ to the list and otherwise behaves as π satisfies $V^{\pi^{+w_0}} \geq V^{\pi}$.

Write $g = u_{\text{TP}} - u^{+}_{\text{IND}}$ and $g^{-} = u_{\text{TN}} - u^{-}_{\text{IND}}$; both are positive by the ordering of Appendix B.2.

Proposition 2. *Under* (H0)–(H2) *and a noiseless channel, every open-loop policy π satisfies*

$$V^{\star} - V^{\pi} \;\geq\; \gamma \;:=\; \min\big\{\, p\,(\lambda c(w') - k' g^{-}),\;\; q\,(g - \lambda c(w'))\,\big\} \;-\; e\,\Delta.$$

Proof. It suffices to exhibit, for each open-loop π, a policy π' with $V^{\pi'} - V^{\pi} \geq \gamma$, since $V^{\star} \geq V^{\pi'}$. We split on whether the list of π contains w_0 and w'. Throughout, "the reassembled history" means the history π would have recorded: under a noiseless channel every observation is a function of the state and the action, so a policy that has taken the same actions in a different order can write down π's history verbatim, and any report π writes is available to it.

Case A: the list contains w_0 and w'. Let π_A execute the same list with $\textsc{scan}(w_0)$ moved to the front. When it reaches $\textsc{scan}(w')$ it reads the cue: if the cue says no-branch it skips w' and every work-up on w'; otherwise it scans as scheduled. At the end it writes the following report. If it did not skip, the report π writes on the reassembled history. If it skipped, it forms the counterfactual history in which w' was scanned, was open, and showed no sign on any finding in F', with the null work-up observations that state determines; it takes the report π writes on that history and overwrites the verdict on every finding in F' with INDETERMINATE.

On the event $\{$cue says branch$\}$, π_A takes the actions of π and writes π's report; costs and utilities agree, so the returns are equal. On W the counterfactual is exactly what π observed, so π_A's report agrees with π's off F'; the scanned sets differ only by w', which by (H0) and the definition of F' reaches none of those findings, so their utilities agree. On each $j \in F'$ the finding is absent, so π earns at most u_{TN} and π_A earns u^{-}_{IND}: a loss of at most g^{-} per finding, $k' g^{-}$ in all. Against this π_A saves $\lambda c(w')$ and the cost of any work-up on w'. So $R^{\pi_A} - R^{\pi} \geq \lambda c(w') - k' g^{-}$ on W. On the remaining event, $\{$no-branch$\} \setminus W$, of probability e_W, π_A has skipped w' on a state where the counterfactual may be wrong; we bound the return difference by $-\Delta$. Taking expectations,

$$V^{\pi_A} - V^{\pi} \;\geq\; p\,(\lambda c(w') - k' g^{-}) - e_W \Delta.$$

Case B: the list contains w_0 but not w'. Let π_B execute the same list with $\textsc{scan}(w_0)$ moved to the front, and, after the last scheduled action, scan w' if and only if the cue said branch; then it reports. Its report is the report π writes on the reassembled history of π's own actions, with one change: if it scanned w' and saw the sign of i, it sets the verdict on i to PRESENT.

On $\{$cue says no-branch$\}$ the returns are equal. On U, π_B scans w', which is open for i, and under a noiseless channel sees the sign; its PRESENT on i is licensed, because w' is adequate and scanned, and earns u_{TP}. The policy π has scanned no window adequate for i, since only w' is: an ABSENT is a false negative (or an unsupported one), a PRESENT is priced as a false positive by Appendix B.2, and its best verdict is INDETERMINATE, worth u^{+}_{IND}. The other findings' verdicts agree, and adding w' to the scanned set can only license, never unlicense, a verdict, so their utilities are no lower for π_B. Hence $R^{\pi_B} - R^{\pi} \geq g - \lambda c(w')$ on U. On $\{$branch$\} \setminus U$, of probability e_U, π_B pays $\lambda c(w')$ for a scan whose observation it does not use, and the same licensing argument shows it loses nothing else; so the difference is at least $-\lambda c(w') \geq -\Delta$. Taking expectations,

$$V^{\pi_B} - V^{\pi} \;\geq\; q\,(g - \lambda c(w')) - e_U \Delta.$$

Case C: the list omits w_0. By (H2), π^{+w_0} is an open-loop policy with $V^{\pi^{+w_0}} \geq V^{\pi}$ whose list contains w_0, so Case A or Case B applies to it and yields π' with $V^{\pi'} - V^{\pi^{+w_0}}$ bounded as above. Chaining the two inequalities gives the same bound against π.

In every case $V^{\pi'} - V^{\pi} \geq \min\{p(\lambda c(w') - k' g^{-}),\; q(g - \lambda c(w'))\} - \max\{e_U, e_W\}\Delta \geq \gamma$. □

Remark 4 (reading the hypotheses). (H1) is the formal content of "a branch point": the state decides whether w' is worth visiting, and an earlier window reveals which. The bound is positive

when both branches carry weight. Specifically, $\lambda c(w') > k' g^-$ makes an empty scan a net loss after accounting for the negatives it licenses, while $g > \lambda c(w')$ makes a scan that reveals the finding a net gain. The cue must also be accurate enough that $e\Delta$ does not exceed the smaller term. With an exact cue ($e = 0$), the bound is $\Theta(\min\{p, q\})$ and Remark 2 is the special case $k' = 1$, $g^- = 0$, $p = 1-\delta$, $q = \delta$. (H2) is the assumption every protocol already makes about its first station; without it a list that skips the cue window could avoid the charge only by forgoing whatever w_0 shows. Nothing in the proof uses which station list π follows, which is why the bound holds for the published list as well as for the oracle protocol of Remark 3.

Remark 5 (a noisy channel). Under (C3$_\beta$) with the coupling of Appendix B.3 modified to draw one noise variable per window rather than per step, so that reordering leaves observations unchanged, the same constructions go through with two changes. In Case B the look at w' on U may depart with probability at most β, in which case π_B keeps INDETERMINATE and has paid $\lambda c(w')$ for nothing, so its gain on U becomes $(1-\beta)\, g - \lambda c(w')$. In Case A a plane failure at w' on W makes π's actual observation differ from the counterfactual, an event of probability at most β on which we bound the difference by $-\Delta$. If the cue itself can be missed, a departure at w_0 misroutes the competitor on an event of probability at most β, again bounded by $-\Delta$. So each term of γ loses at most $\beta\Delta$, which is the statement made in § 3.1.

B.5 Divergence by cause

Two separable mechanisms can make an open-loop and a belief-conditioned policy choose different actions, and they belong to different conditions. A *miss* ($\beta_i > 0$) or a transient no-plane return at an otherwise adequate window is observation noise and therefore a failure of (C3). By contrast, a *patient-specific closed window* ($\omega_w = 0$ in the realized Ω although the modal state has it open) is variation in the latent state and therefore a failure of (C2) on the Ω component. A window that is already closed in the modal state violates neither condition: the noiseless trajectory simply avoids it. A miss can be addressed by looking again, and a policy that re-looks on an empty window and abstains when its looks still leave a miss too likely is closed loop on its own record but carries no belief. It decides *how many* times to look, but not *where*. A patient-specific closed window or branch point instead requires a belief about which window is worth the next look.

We report the divergent fraction by cause because a family may require repeat looks without requiring window selection. ρ_w is the share of cases with a present finding that no adequate protocol window can license but another window can. ρ_Ω is the share in which a protocol window is closed on this patient, so the protocol's negative there is unsupported and a substitute view is needed. ρ_{flow} is the share needing color Doppler somewhere, a layer decision that a protocol can make by rule. We exclude cost-only divergence, where the optimum skips an empty protocol window solely to save probe time.

Table 4: Divergent-case fractions by examination family and cause. The report-grounded rows use their public test splits; focused emergency pools its $60/60$ synthetic train/test cases because its scenario specification defines both splits.

Family	n	ρ_w	ρ_Ω	ρ_{flow}	Unreach.
Thyroid	412	0.01	0.02	0.61	0.02
Breast	395	0.03	0.06	0.26	0.06
Liver	168	0.00	0.62	0.03	0.08
Focused emergency	120	0.51	0.72	0.00	0.18

Budget matching guards the predictions of § 3.1 from the cost side: a belief-based policy can win on *cost* alone by skipping scans whose EVOI does not repay $\lambda c(w)$, so we also score each policy under the other's realized budget on the same case.

B.6 Case generation and examination families

Case generation. A case consists of a state (y, ω) and an indication. The three *report-grounded* families, thyroid, breast, and liver, take y from structured public ultrasound reports. The indication is generated from the hidden case, and ω is sampled by a seeded probability model rather than inferred

from the report. Thus the findings are report-grounded, while the presentation, acoustic obstacles, and unobserved counterfactual trajectories are environment-generated.

The *focused-emergency* family (eFAST/RUSH) has no report corpus and is fully synthetic. Its sampling specification defines the scenarios, from which the generator samples indications, correlated findings, and access obstacles using fixed seeds. The sampled (y, ω) is rendered into the same structured case format so that the indexer, simulator, and checker run unchanged on both kinds of family. These synthetic cases are a benchmark stress test, not an estimate of clinical prevalence or performance.

Examination families. SONOARENA instantiates four families: thyroid, breast, liver, and focused emergency (eFAST/RUSH). They span the acquisition regimes analyzed in Propositions 1–2. A thyroid study is protocol-complete: its station list covers every window the report will mention, the budget affords a full sweep, and an open-loop policy is already close to optimal. In a focused emergency study, routine views account for barely half the windows the indication can require, the rest (cardiac substitution views, inferior vena cava, lower-limb veins, aorta) are opened only when the belief warrants it, and several fail outright. Table 4 reports the measured divergent fraction per family: on thyroid $\rho_w = 0.01$ and $\rho_\Omega = 0.02$, with the decision that exists there being about Doppler on a present lesion ($\rho_{\text{flow}} = 0.61$); on liver $\rho_\Omega = 0.62$ and $\rho_w = 0$; on focused emergency $\rho_w = 0.51$ and $\rho_\Omega = 0.72$, with 18% of cases carrying a present finding that no adequate window can show.

B.7 THE THREE ARTIFACTS THE CHECKER READS

The checker reads three artifacts separately against the same hidden state.

1. **Trajectory.** The recorded action path: which windows were attempted, which returned a usable plane, what they revealed, and what they cost. It determines coverage, acquisition time, and whether a claim is licensed. An ABSENT verdict is supported exactly when some acquired window in $\Omega_{\cdot i}$ returned a usable plane.
2. **Structured findings.** One verdict per diagnostic question from $\{\text{PRESENT}, \text{ABSENT}, \text{INDETERMINATE}\}$, paired with its finding and window. This axis yields return, regret, abstention, and unsupported positive and negative rates without language understanding. Omitted questions are scored as INDETERMINATE; a false negative following an adequate view is an incorrect but licensed miss.
3. **Free-text report.** The narrative a clinician would read. We split it into clauses deterministically, then ask a frozen LLM attributor at temperature zero to assign each clause a window and polarity. We use `gpt-4o-mini` by default and `claude-sonnet-4-6` for reports produced by OpenAI-backed agents, so the same LLM never both generates and attributes a report. The attributor sees only the report and the window vocabulary, without (y, Ω), the trajectory, or the structured verdicts. Once attributed, a deterministic set-membership check yields text-based unsupported negatives, contradictions, and undeclared claims. Answers are cached with model and prompt provenance. A refused report is sent to the other provider; failed calls do not fall back to a rule parser.

Scoring the three artifacts separately exposes disagreements that a single axis misses. A structured INDETERMINATE paired with a narrative "no abnormality" is invisible to either axis alone, and is exactly the failure a report reader would act on. Two separate 0–100 monitors accompany them: documentation quality, scored on the clinical QC checklist of Appendix C, and agreement of the ultrasound impression with the ground truth. Neither enters the terminal return or the learning reward.

Attributor calibration. Calibration examples are generated by construction. To estimate precision, the report is formed only from fragments actually revealed to the player, so any unsupported claim flagged by the pipeline is a false positive. For recall, we remove one window from the recorded trajectory while leaving its revealed fragments in the report; every clause indexed to that window should then be flagged. This protocol isolates the model-dependent attribution step; all subsequent adjudication remains deterministic.

B.8 TASK-TRAINED POLICY DETAILS

Common policy configuration. All task-trained policies use `Qwen3-8B` with thinking disabled, bfloat16 weights, gradient checkpointing, and a maximum sequence length of 3072. LoRA uses rank 32, scale 64, dropout 0.05, and targets the `q`, `k`, `v`, and `o` attention projections and the `gate`, `up`, and `down` MLP projections. Episodes are capped at 28 actions. If a sequence is too long, we truncate the prompt head but never the policy response.

SFT. We sample 800 training cases across the four families. Two non-privileged teachers, the belief-based HYDRA planner and the re-look-and-abstain protocol, each generate two stochastic episodes per case. We retain the highest-return trajectory for each case and deduplicate identical prompt–response turns. Training uses one epoch of response-only cross-entropy, AdamW with learning rate 10^{-4}, micro-batch size 1 and gradient accumulation 8, a 5% validation split, 3% warm-up followed by cosine decay, and gradient-norm clipping at 1.

GRPO. We train for 50 iterations, sampling 8 cases and 8 rollouts per case at each iteration. Rollouts use temperature 1.0 and top-$p = 0.95$; all members of a case group share the same simulator seed, while the seed changes across iterations. We use one optimization epoch of token-level clipped GRPO with clip 0.2, AdamW at 10^{-5}, micro-batch size 2, gradient-norm clipping at 1, and no explicit KL penalty. The two initializations are the base model and the SFT adapter. Evaluation uses greedy decoding (temperature 0). The free-text quality score is logged only as a reward-hacking monitor; the reward is the deterministic structured return. The case-specific analytical ceiling is used only to report regret and does not enter the GRPO advantage.

B.9 HARNESS FOR THE GENERAL-PURPOSE LLM-BASED POLICIES

The general-purpose LLM-based policies in Table 1 use the same ReAct agent harness and closed-loop interface. Prompts, observations, tool semantics, and the probe-time budget are fixed across rows; decoding exceptions are listed below. An episode contains a sequence of *scan turns* followed by one *report turn*.

Scan turn. A fixed system prompt, identical for every model, describes the action set (`scan`, `cdfi`, `b`, `detail`, `measure`, `capture`, and `finish`) and their preconditions (a `detail` needs a lesion already seen, a `capture` a scanned window in the current mode). The user message is a JSON object with the chief complaint, the examination family, the current mode, the remaining time budget, the window menu with each window's probe cost, the distractor windows, and the last 12 lines of the observation log returned by the environment. The model must answer with a single JSON object {`"action", "target"`}; the prompt asks for no explanation and no reasoning trace is requested or scored (models that expose a thinking mode, `deepseek-v4-pro` and the `qwen3.x` family, run with it disabled). The action is parsed literally: `finish` ends the examination; `cdfi`/`b` switch mode; `detail`, `measure` and `capture` act on the named window; any other action with a target is a scan of that window in the current mode; an action without a target ends the examination. A response that is not valid JSON is sent back with the parse error for repair, up to four times; a turn that still fails is recorded as an agent error and the episode is re-run. Episodes are capped at 40 actions and by the environment's time budget; no policy reached the action cap (mean 6–18 actions per examination, maximum 37).

Report turn. A second fixed system prompt asks for a JSON report with a free-text description, an impression, and a list of structured findings, each carrying a window, a verdict in {PRESENT, ABSENT, INDETERMINATE} and the log line it rests on; it states the licensing rule in words (a window that was not scanned must be reported INDETERMINATE, never ABSENT). The user message contains the family, the chief complaint, the window list, the sets of windows scanned in B-mode and in color Doppler, and the complete observation log. The checker reads the verdicts (structured axis) and the description (prose axis, § 2.3). The impression does not enter the return; it is used only by the diagnostic-agreement monitor.

Full prompts used in evaluation. The boxes below give complete English translations of the deployed system messages. JSON keys, allowed values, constraints, and the dynamic user-message

schemas are unchanged. Angle-bracketed values are filled by the environment at each turn. These are the same prompts used by the API ReAct policies and the local `Qwen3` policies.

Scan-turn prompt

System message.

```
You are an ultrasound resident conducting an examination. Based on
the findings acquired so far and the remaining budget, choose the next
step.
Output JSON only:
{"action": "scan" | "cdfi" | "b" | "detail" | "measure" | "capture" |
"finish",
 "target": "<name>"}
Rules:
- scan: first survey of a window; it reveals only the overview or
whether a lesion is present.
- detail: inspect morphology after a lesion has been found (requires
a prior scan that found the lesion).
- measure: an additional measurement step (lesion dimensions or
organ dimensions; organ measurement requires that window to have been
scanned).
- capture: save the current virtual image (requires that window to
have been scanned in the current mode).
The target is required for scan, detail, measure, and capture. Do not
explain.
```

User message at each scan turn.

```
{
  "chief_complaint": <string>,
  "organ": <examination-family key>,
  "mode": "b" | "cdfi",
  "budget_left": <remaining probe-time units>,
  "menu": ["<key>/<name>(cost=<value>)", ...],
  "distractors": ["<key>/<name>", ...],
  "log": [<last 12 observation-log lines>]
}
```

Report-turn prompt

System message.

```
You are an ultrasound resident. Write the report from the examination
log.
Output JSON only:
{"description": "<ultrasound findings>", "impression": "<ultrasound
impression>", "findings": [...]}
If quotation marks are needed inside a JSON string, escape the double
quotation marks.
Give one structured judgment per finding:
{"finding": "<finding name>", "region": "<window key>",
 "verdict": "present" | "absent" | "indeterminate",
 "is_flow": false, "evidence": "<supporting log entry>"}
- present: actually observed.
- absent: the window was scanned, the image was adequate for
judgment, and the finding was confirmed absent.
- indeterminate: the window was not scanned or the image was
inadequate. If a window was not scanned, the verdict must be
indeterminate, never absent.
The structured findings and description must agree. Prefer
indeterminate to an unsupported absent finding.
Use standard ultrasound terminology and a standard report-template
structure (size, shape, capsule, parenchyma, lesion features, and
CDFI).
```

```
The impression must include anatomical location, diagnostic
characterization (e.g., TI-RADS or disease name), and follow-up,
biopsy, or other recommendations.
```

User message at the report turn.

```
{
  "organ": <examination name>,
  "chief_complaint": <string>,
  "regions": [{"key": <key>, "name": <name>}, ...],
  "scanned": [<B-mode window keys>],
  "cdfi_scanned": [<CDFI window keys>],
  "findings_log": [<complete observation log>]
}
```

Decoding and providers. API models are called through an OpenAI-compatible endpoint in JSON mode at temperature 0.2 for both turns, except `kimi-k3`, whose provider accepts only temperature 1, and `claude-opus-5`, which runs through the vendor's headless CLI with the provider's default sampling. When a provider's content filter refuses a case (`gpt-5.5` and `gpt-5.4-mini` on part of the focused-emergency and thyroid packs), the episode is excluded and the cell is scored on the accepted cases with the abstention floor recomputed on the same cases; `grok-4.6` and `glm-5.3` were run on the first 25 cases of each family.

Open-weight and task-trained policies. The `Qwen3` rows and all `Qwen3-8B` variants use the same two system prompts and the same JSON fields through a local `transformers` harness (bfloat16, thinking disabled). The probe-time budget is unchanged, but they differ from the API harness in three settings only: greedy decoding instead of temperature 0.2; a cap of 28 actions instead of 40 (reached in 1–6 of 360 examinations per row); and generation limits of 96 new tokens per scan turn and 2048 per report. A malformed action is treated like an unrecognized target and costs the step without revealing anything. Both harnesses show the same 12 log lines and neither adds a summary of the windows already visited; the trained policies were trained under exactly this interface.

B.10 The analytical ceiling

The SonoScore normaliser is an optimistic deterministic relaxation, not the executable clairvoyant policy reported in Table 1. For every selected adequate window, the relaxation assumes that the first acquisition returns a usable plane and reveals every finding available through that window. It pays the real first-look cost once, with no stochastic miss, transient plane failure, or retry cost. Each finding is then either covered by some selected window and reported correctly, or left INDETERMINATE. The resulting optimization is a prize-collecting weighted set cover over $\{w : \Omega_{wi} = 1\}$, solved exactly by enumeration at our scale ($|\mathcal{W}| \le 11$, $m \le 12$). If a finding has no adequate window, it is necessarily left indeterminate, so $\bar{V}$ is defined on every case.

For any realized executable trajectory, replace every window it opened by the relaxation's successful one-look version. This weakly increases obtainable report utility and weakly decreases acquisition cost. The relaxation may then choose an even better window subset. Hence $R(\tau) \le \bar{V}$ trajectory by trajectory, and therefore $V^{\pi} \le V^{\star} \le \bar{V}$. The executable clairvoyant reads (y, Ω) but still acts through the stochastic gateway; it is therefore a privileged policy below this analytical ceiling rather than the ceiling itself.

B.11 Where Hydra and the clairvoyant sit relative to $V^{\star}$

Every policy π, Hydra included, satisfies

$$V^{\pi} \;\le\; V^{\star} \;\le\; \bar{V},$$

so the regret we report splits into two non-negative parts,

$$\underbrace{\bar{V} - V^{\pi}}_{\text{reported regret}} \;=\; \underbrace{\bar{V} - V^{\star}}_{\text{relaxation gap}} \;+\; \underbrace{V^{\star} - V^{\pi}}_{\text{optimality gap of } \pi} .$$

The expected regret of a policy upper-bounds both its optimality gap and the looseness of $\bar{V}$. The first component is common to all policies, so the difference in expected regret between two policies equals the difference in their values. Empirical tables estimate these quantities by averaging episode-level values. What remains is to say how large each expected component is. We can do so in closed form in a canonical case, and structurally beyond it. Throughout this subsection "HYDRA" means the controller of § 3.2 run with the true prior and the true miss rates; what a wrong prior costs is an empirical matter, measured by the −LLM-PRIOR ablation.

What is optimal by construction. Lemma 1 (report). *Given the posterior b and the record of usable planes, the report of § 3.2 maximizes expected utility among reports that assert* PRESENT *only on an observed sign.* The utility is a sum over findings, so the verdicts can be chosen one at a time. A sign gives $b_i = 1$ under $\alpha = 0$ and u_{TP} is the largest entry. Without a sign, ABSENT has expected utility $A(b_i) = b_i u_{\text{FN}} + (1 - b_i)u_{\text{TN}}$ when a usable plane from a window that reaches finding i is on record, and at most INDETERMINATE's otherwise by the constraint on ϕ (Appendix B.2); INDETERMINATE has $I(b_i) = b_i u^{+}_{\text{IND}} + (1 - b_i)u^{-}_{\text{IND}}$. Both are linear in b_i and cross at $p^{\dagger}$, which is the threshold the report uses. The acquisition rule at $M = 1$ is, by its definition, optimal among policies that take one more action and then report; the implemented rule scores plans of up to $M = 4$ looks and is characterized below. Neither fact makes the policy optimal; the rest of this subsection says when it is and what it loses when it is not.

The canonical case. One finding with prior b; one window that reaches it and is known to be open; each look costs c and misses a present finding with probability β. Write $G = u_{\text{TP}} - u_{\text{FN}}$, $g = u_{\text{TP}} - u^{+}_{\text{IND}}$, $g^{-} = u_{\text{TN}} - u^{-}_{\text{IND}}$, and assume a look is worth taking for a reader who knows the truth: $g \geq \lambda c/(1 - \beta)$ and $g^{-} \geq \lambda c$. A transient plane failure with probability p_f leaves the belief unchanged and is absorbed by replacing c with $c/(1 - p_f)$ for every policy alike, so we set $p_f = 0$.

When the finding is absent no look shows a sign, so along that path a policy's decisions depend only on how many looks it has taken; and once a sign appears the finding is settled and further looks only cost. Every undominated policy is therefore a pair (n, r): look until the first sign or until n looks, and report $r \in \{\text{ABSENT}, \text{INDETERMINATE}\}$ if none appeared. Its value is

$$V(n, r) \;=\; b\Big[(1 - \beta^n)\, u_{\text{TP}} + \beta^n u^{+}_{r} - \lambda c\, \tfrac{1-\beta^n}{1-\beta}\Big] \;+\; (1 - b)\big[u^{-}_{r} - n\lambda c\big],$$

with $(u^{+}_{r}, u^{-}_{r}) = (u_{\text{FN}}, u_{\text{TN}})$ for ABSENT, which needs $n \geq 1$ to be licensed, and $(u^{+}_{\text{IND}}, u^{-}_{\text{IND}})$ for INDETERMINATE. So $V^{\star} = \max_{n,r} V(n, r)$ is computable by inspection. The analytical relaxation obtains the sign in one look on either state, so $\bar{V} = b\,(u_{\text{TP}} - \lambda c) + (1 - b)(u_{\text{TN}} - \lambda c)$.

Proposition 3 (the two gaps in the canonical case). *(a) Looseness of the bound. For every $n \geq 1$,*

$$\bar{V} - V(n, \text{ABSENT}) \;=\; \underbrace{b\Big[\beta^n G + \lambda c\Big(\tfrac{1-\beta^n}{1-\beta} - 1\Big)\Big]}_{\text{residual miss and retry cost}} \;+\; \underbrace{(1 - b)(n - 1)\,\lambda c}_{\text{confirmatory looks}},$$

and for every $n \geq 0$, $\bar{V} - V(n, \text{INDETERMINATE}) = b[\beta^n g + \lambda c((1 - \beta^n)/(1 - \beta) - 1)] + (1 - b)[g^{-} + (n - 1)\lambda c]$. Hence $\bar{V} - V^{\star}$ is the smallest of these over (n, r), and it is strictly positive whenever $0 < b < 1$, $\beta > 0$ and the standing assumptions hold strictly.

(b) Optimality of the myopic rule. If $p^{\dagger}(1 - \beta)\, g > \lambda c$, then HYDRA *attains $V^{\star}$: it stops at*

$$n^{\star} \;=\; \min\Big\{n \geq n_0 :\; b\,\beta^n\big[(1 - \beta)G - \lambda c\big] \leq (1 - b)\,\lambda c\Big\},$$

where $n_0 \geq 1$ is the first n at which the no-sign posterior $b_n = b\beta^n/(b\beta^n + 1 - b)$ is below $p^{\dagger}$, and reports ABSENT*; or it takes no look at all when the first look has non-positive value, which is then also optimal.*

Proof. (a) Subtract. When the finding is present the policy stops short with probability β^n, forgoing $u_{\text{TP}} - u^{+}_{r}$, and also pays retry costs that the one-look relaxation removes; when it is absent the policy pays for n looks where the relaxation pays for one, and under INDETERMINATE also forgoes g^{-}. Each bracket is non-negative by the standing assumption, and strictly positive when it holds strictly; with $0 < b < 1$ and $\beta > 0$ every expression is then positive.

(b) Write $P_n = b\beta^n + 1 - b$ for the probability of no sign in n looks. By Bayes' rule and the linearity of A and I,

$$\begin{aligned} &V(n{+}1, \text{ABSENT}) - V(n, \text{ABSENT}) \\ &\quad = b\beta^n \big[(1-\beta)G - \lambda c\big] - (1-b)\lambda c \\ &\quad = P_n \cdot \text{EVOI}(b_n), \qquad n \geq 1, \end{aligned}$$

and the same identity holds with g for G between consecutive INDETERMINATE policies. Both increments decrease in n, because $(1-\beta)G > (1-\beta)g \geq \lambda c$. While $b_n \geq p^\dagger$ the report would be INDETERMINATE and $\text{EVOI}(b_n) \geq b_n(1-\beta)g - \lambda c \geq p^\dagger(1-\beta)g - \lambda c > 0$, the first inequality because the post-look report value is at least I; so HYDRA continues, and so does the optimal policy, since $V(n{+}1, r^\star) \geq V(n{+}1, \text{INDETERMINATE}) > V(n, \text{INDETERMINATE})$. From n_0 on the report is ABSENT, the increments are decreasing, and the maximiser is the first n at which the increment is non-positive, which is where $\text{EVOI}(b_n) \leq 0$ first holds. If $b < p^\dagger$ at the outset, the first look is worth $[\, b(1-\beta)G - \lambda c \,] + [\, A(b) - I(b) \,]$, the second bracket being the value of licensing the negative; if this is non-positive then $b(1-\beta)G < \lambda c$, every later increment of either kind is negative, and not looking is optimal. A budget that binds truncates both policies at the same n. □

Remark 6 (reading Proposition 3). Part (a) says what the analytical ceiling credits that no executable policy can generally realize: it removes the chance β^n that a present finding is still unseen when a rational policy stops and all retry costs, including the $n-1$ looks a rational policy spends on a negative patient to push its belief below $p^\dagger$. At low prevalence the second term carries weight $1-b$ against $b\beta^n$ for the first, so unless a miss is priced very high a large share of $\bar{V} - V^\star$ is the cost of *earning* a negative rather than of missing a positive. Part (b) says the myopic rule is exact when a single look at the abstention threshold pays for itself through detection alone. When that fails, the rule can stop at some $n < n_0$ with INDETERMINATE although n_0 looks would have bought a licensed negative worth more than they cost. This is the familiar non-concavity of the value of information (Radner & Stiglitz, 1984), and the loss, $\max_n V(n, r_n^\star) - V(n_{\text{H}}, \text{INDETERMINATE})$, is read off the closed form. Evaluating "up to m further looks, stopping at the first sign" for every m within budget would compute $\max_n V(n, r_n^\star)$ directly and remove the condition. The implemented rule takes this form but not this value, for the reason given next, so Proposition 3(b) is stated for $M = 1$.

The implemented rule. HYDRA does not evaluate a single look. For each window it scores the plan "look up to m times, stopping at the first sign" for every $m \leq M$ and keeps the largest value, executes one look at the best window, and re-plans; $M = 4$, the same constant that caps re-looks at one window. The utility side of this score is exactly the $V(n, r_n^\star)$ of the canonical case: with per-look reveal probability $q = 1 - \beta$, a sign arrives within m looks with probability $b(1-\beta^m)$ and is worth u_{TP}, and otherwise the posterior $b\beta^m/(b\beta^m + 1 - b)$ is reported optimally with a negative licensed. The cost side is not: the plan is charged λmc on every branch, although a plan that stops at the first sign pays $\lambda c(1-\beta^m)/(1-\beta)$ in expectation when the finding is present. The implemented score is therefore a lower bound on $V(m, r_m^\star) - U^*(b)$, tight when $\beta \to 1$ or $m = 1$ and loose by $b\lambda c\,[\, m - (1-\beta^m)/(1-\beta) \,]$ in general. The rule can stop while a longer plan would still repay itself, so it does not inherit the exactness the previous paragraph describes, and Proposition 3(b) is not extended by it. Under deterministic observations $q = 1$, the maximiser is $m = 1$ for every window, so the rule degenerates to one-step EVOI exactly; the horizon matters only in the stochastic regime the evaluation uses.

Beyond the canonical case. The EVOI rule is not generally globally optimal at any horizon. Threshold crossing can make a sequence of negative looks worthwhile even when its first look is not (Remark 6); scoring plans of up to M looks answers this *within* a window, but not across windows, where the rule still compares single acquisitions. Under uncertain access, an initial usable plane can raise the value of scanning the same window again; acquisitions may also complement one another or jointly inform several findings. The factored belief uses correlations between findings only to the extent that the prior encodes them. These limitations contribute to $V^\star - V^{\text{HYDRA}}$, rather than to the looseness $\bar{V} - V^\star$ of the analytical ceiling.

The bound under uncertain access. With several windows of unknown access, $\bar{V} - V^\star$ gains a term the canonical case lacks: the clairvoyant goes straight to the cheapest adequate cover, whereas

any policy pays $\lambda c(w)$ for each closed window it tries on the way to an open one. We do not bound this term in closed form. Its frequency is what ρ_Ω records (Table 4), so it should be largest in the families where windows close most often. Together with Proposition 3(a) it gives the reading of regret we recommend: $\bar{V}$ is a ceiling on $V^\star$ that overstates it by the cost of confirmatory looks, the residual miss, and the search for an open window, and a policy's regret should be judged against that overstatement rather than against zero.

B.12 A WORKED EPISODE: THE BELIEF STATE STEP BY STEP

The belief is a table with one row per question: the probability $b_{t,i}$ that finding i is present, the agent's own record for the windows that can show it (looks made, usable planes obtained), and a status (open, closed, settled). Below is one focused-emergency case under the default parameters: prior from the train-split base rates, reveal rate $q = 0.5$ per usable look, and the absent threshold $p^\dagger = 0.135$ solved from the utility matrix (§ 3.2). Chief complaint: blunt trauma after a road accident, abdominal pain, hypotension.

Table 5: Initial belief state before any acquisition (excerpt).

Finding i	Windows $A_{\cdot i}$	$b_{0,i}$	Record (looks/usable)	Status
Free fluid, Morison's pouch	morison	0.124	0/0	open
Free fluid, splenorenal	splenorenal	0.050	0/0	open
Free fluid, pelvis	pelvis	0.098	0/0	open
Pericardial effusion	subxiphoid, PLAX, apical	0.126	0/0	open
RV strain	apical, PLAX, subxiphoid	0.176	0/0	open
Pneumothorax, right	lung_r	0.118	0/0	open

$t = 1$**:** SCAN**(morison)** $\rightarrow$ **usable plane, a sign.** The clause "free fluid in Morison's pouch" is returned. Case one: $b_{\text{ruq}} \leftarrow 1$, status settled, and detail/measure on that window are queued for work-up.

$t = 2$**:** SCAN**(subxiphoid)** $\rightarrow$ **no usable plane, cue *closed*.** The cue carries the report's wording ("subxiphoid view obscured by costal margin and bowel gas"). Case three, $\Lambda = 1$: $b_{\text{pericardial}}$ stays at 0.126. The window enters the closed set; the hypothesis it carried is handed to PLAX and apical, whose beliefs are raised to $\max(\cdot, 0.126)$.

$t = 3$**:** SCAN**(PLAX)** $\rightarrow$ **usable plane, nothing.** Case two for every finding PLAX can show. Pericardial effusion: odds $0.144 \rightarrow 0.072$, $b = 0.067$. RV strain: odds $0.214 \rightarrow 0.107$, $b = 0.097$. Record for PLAX: 1/1.

$t = 4$**:** SCAN**(pelvis)** $\rightarrow$ **no usable plane, cue *transient*.** Case three again: $b_{\text{pelvis}} = 0.098$ unchanged; record 1/0; the window stays open. At prior 0.098 and $u_{\text{FN}} = -19$, the value of one more look exceeds its cost.

$t = 5$**:** SCAN**(pelvis)** $\rightarrow$ **usable plane, nothing.** b_{pelvis}: odds $0.109 \rightarrow 0.054$, $b = 0.052$; record 2/1.

$t = 6$**:** SCAN**(lung_r)** $\rightarrow$ **usable plane, nothing.** $b_{\text{ptx,r}} = 0.063$; record 1/1.

$t = 7$**: stop.** No remaining open window has positive acquisition value under the controller's stopping rule; the agent completes the queued work-up and reports.

Table 6: Report synthesis from the posterior and acquisition record at $t = 7$.

Finding	$b_{7,i}$	Record	Verdict	Why
Free fluid, Morison's pouch	1	1/1	PRESENT	seen
Pericardial effusion	0.067	subx. 1/0 (closed); PLAX 1/1	ABSENT	$b < p^\dagger$ and PLAX licenses it
RV strain	0.097	PLAX 1/1	ABSENT	likewise
Free fluid, pelvis	0.052	2/1	ABSENT	second look obtained a plane
Pneumothorax, right	0.063	1/1	ABSENT	licensed
Free fluid, splenorenal	0.050	0/0	INDETERMINATE	$b < p^\dagger$, but no window was ever opened

The splenorenal row illustrates the distinction. Its belief is *lower* than the pericardial one, yet it cannot be called ABSENT: the belief decides whether a negative is worth asserting, the record decides

whether it may be. On the same case the fixed protocol writes “no pericardial effusion” through a closed subxiphoid window and “no pelvic fluid” after a single failed plane. These are unsupported negatives, even though one happens to be correct. The protocol never opens the parasternal view.

C ULTRASOUND EXAMINATION QUALITY RUBRIC

Table 7 translates the 100-point checklist used for the quality check (QC) column of Table 2. The simulator cannot observe HIS demographics, gain and scale settings, or handwriting; these items are marked not applicable and receive full credit. Doppler items are scored by whether the corresponding capture succeeds. QC therefore measures stored-image completeness and report structure, not image physics. It is a monitoring metric and does not enter the return or training reward (§ 4).

Table 7: Clinical ultrasound quality checklist underlying the QC column of Table 2.

Category	Criterion	Points	Deduction rules	Responsible	Score
General information (5 pts)	Patient name, sex, age, medical record number, requesting department, ward, bed number, examination item, examination equipment	5	0.5 pt deducted for each item that does not match the actual record	Physician, assistant	
Ultrasound images (50 pts)	The image carries the patient's name or identification number	2	2 pts deducted if absent	Physician	
	The examination is performed according to the requested items; the stored image set is complete and reflects the features of the disease	10	2–8 pts deducted for missing images	Physician	
	Focal lesions are documented in $\geq$2 planes of different orientation, clearly showing the relationship with surrounding structures and the grayscale and color Doppler features	10	4 pts deducted for a single plane only; 4 pts if the relationship with surrounding structures is unclear; 2 pts if no color Doppler image	Physician	
	Images are annotated with a body marker or a text annotation	4	2 pts deducted per image lacking a body marker or text annotation; 1 pt if the transducer position is incorrect	Physician	
	Size and depth of the organ or target lesion are displayed, with appropriate focus	4	2 pts deducted per substandard image	Physician	
	Grayscale images are clear with appropriate gain	4	2 pts deducted per substandard image	Physician	
	Color Doppler gain, scale, and size and angle of the sample box comply with the standard	4	2 pts deducted per substandard image	Physician	
	Spectral Doppler gain, scale, sample volume, and angle comply with the standard	4	2 pts deducted per substandard image	Physician	
	Measurements comply with the standard	4		Physician	
	The images selected for the report are relevant, representative, and meaningful for the present examination	4	2–4 pts deducted for inappropriate image selection	Physician, assistant	

Category	Criterion	Points	Deduction rules	Responsible	Score
Report documentation (45 pts)	All items specified on the request form are described, with a targeted account of the question raised on the request form	5	5 pts deducted for each missing item	Assistant	
	The provincial standard ultrasound report template is used	5	Full 5 pts if used throughout; points deducted as appropriate if used only in part	Physician, assistant	
	Standardized ultrasound terminology is used	5	2 pts deducted per non-standard instance	Physician, assistant	
	Features of space-occupying lesions are described according to the standard	5	2 pts deducted per non-standard description	Physician, assistant	
	The organ, region, and orientation of the lesion (left/right, upper/lower, anterior/posterior) are stated correctly	5	5 pts deducted per error	Assistant	
	Units and numerical data are correct	5	5 pts deducted per error found	Assistant	
	The printed report contains no misspellings; handwritten ultrasound report forms are neat, legible, free of misspellings, and free of alterations	5	5 pts deducted per error found	Physician	
	The ultrasound impression corresponds to the description	5	5 pts deducted per mismatch	Assistant	
	The ultrasound impression gives an anatomical localization and a diagnostic characterization, together with a reasonable recommendation	5	2 pts deducted if localization is missing; 2 pts if characterization is missing; 1 pt if a recommendation is missing	Physician	
Total 100 pts		Reviewing expert			Total score

Note: Excellent $\geq$95 pts; Pass 85–94 pts; Fail $<$85 pts.

D ADDITIONAL RESULTS

This section provides error analyses, policy ablations, and test-time parameter shifts. Analyses that use a different sampling protocol from the main comparison state it in their captions.

Table 8: Negative-reporting errors by cause, per examination, in a separate five-seed diagnostic sweep. Task-trained policies are also pooled over training seeds.

	Unsupported		Licensed
Policy	Never scanned	No usable plane	Missed finding
Fixed protocol	0	1.36	0.42
+ re-look + abstain	0	0	0.07
ReAct (`gpt-4o-mini`)	0.01	0	0.13
HYDRA	0	0	0.20
Zero-shot policy (`Qwen3-8B`)	0.52	0.02	0.09
SFT policy	0	0	0.17
GRPO policy	0.21	0.03	0.10
SFT → GRPO policy	0.06	0.03	0.31
Clairvoyant policy	0	0	0.06

D.1 EVIDENTIAL-LICENSING ABLATION

In a three-seed licensing ablation, removing evidential licensing from the GRPO reward leaves coverage similar (74.1% versus 75.0%), raises unsupported negatives from 0.10 to 0.49 per examination, and lowers SonoScore from 58.9 to 53.7. These experiments suggest that imitation learns re-acquisition and reporting rules more readily than belief-conditioned window selection, while verifier licensing reduces unsupported assertions.

Table 9: Licensing as a verifier-side hard constraint (diagnostic sweep; Overall SonoScore is the unweighted mean across examination families). A programmatic filter rewrites each unsupported ABSENT or PRESENT verdict to INDETERMINATE and leaves the recorded trajectory unchanged.

Policy	UN before ↓	SonoScore	SonoScore, filtered	Δ
Fixed protocol	1.15	26.5	44.5	+18.0
HYDRA without its licensing constraint	0.31	64.0	66.0	+2.0
ReAct `gpt-4o-mini`	0.01	23.7	23.8	+0.0
ReAct `claude-sonnet-4-6`	0.13	46.2	47.7	+1.5
ReAct `gpt-5.5`	0.01	54.0	54.1	+0.1
ReAct `gemini-3.1-pro`	0.01	55.9	56.3	+0.4
ReAct `deepseek-v4-pro`	0.02	32.0	32.3	+0.3
ReAct `claude-opus-5`	0.03	52.7	52.8	+0.2
ReAct `claude-sonnet-5`	0.13	44.1	45.3	+1.3
ReAct `gpt-5.4-mini`	0.12	33.0	34.1	+1.0
ReAct `gemini-3.8-flash`	0.02	52.0	52.5	+0.5
ReAct `grok-4.6`	0.00	51.6	51.6	-0.0
ReAct `glm-5.3`	0.02	49.5	50.2	+0.7
ReAct `qwen3.8-max`	0.04	51.6	51.8	+0.2
ReAct `qwen3.7-plus`	0.28	29.1	32.8	+3.8
ReAct `qwen3.6-plus`	0.28	30.3	33.3	+3.0
`Qwen3-8B` zero-shot	0.26	25.9	28.8	+2.9
`Qwen3-8B` GRPO	0.17	31.2	32.7	+1.5
`Qwen3-8B` SFT → GRPO	0.08	56.3	57.1	+0.8

D.2 Robustness to test-time observation shifts

Table 10: Test-time parameter shifts (SonoScore per family). Every policy is re-run in an environment whose parameters differ from those it was built or trained with, while the policy itself is unchanged: the reveal probability $1-\beta$ is halved (harder) or multiplied by 1.5 (easier), and the probability that an open window returns a usable plane is lowered from 0.9 to 0.7 and 0.5. Belief-based policies retain their original parameters. —: not run; ‡: content-filter subset of Table 1; §: 20 cases of the family, the first ten of the pack plus ten chosen so that the policy's in-distribution SonoScore on the subset equals its full-pack SonoScore, with the floor computed on the same cases.

		Reveal probability		Plane success	
Policy	In-dist.	harder	easier	0.7	0.5
Thyroid					
Fixed protocol	21.7	-0.1	32.3	-3.7	-36.1
Steelman (re-look + abstain)	53.1	34.7	60.2	48.7	39.9
HYDRA-Teacher	59.3	39.9	67.2	58.7	47.5
SFT → GRPO (`Qwen3-8B`, seed 1)	56.9	40.6	65.7	52.3	41.0
ReAct `gpt-5.5`	42.2‡	27.2§	51.0§	47.3§	42.3§
Breast					
Fixed protocol	25.6	17.1	37.1	-4.5	-50.2
Steelman (re-look + abstain)	62.3	58.6	65.6	57.8	49.7
HYDRA-Teacher	57.7	49.7	69.1	56.0	49.5
SFT → GRPO (`Qwen3-8B`, seed 1)	62.6	52.7	69.5	57.7	51.9
ReAct `gpt-5.5`	54.6	61.2§	54.1§	50.4§	59.5§
Liver					
Fixed protocol	54.7	56.6	55.0	25.3	-3.9
Steelman (re-look + abstain)	65.4	67.5	65.0	66.5	64.2
HYDRA-Teacher	92.3	94.6	91.2	93.8	92.1
SFT → GRPO (`Qwen3-8B`, seed 1)	74.2	76.4	71.9	75.9	83.8
ReAct `gpt-5.5`	75.7	—	—	—	—
Focused em.					
Fixed protocol	4.0	1.4	10.4	-13.8	-24.4
Steelman (re-look + abstain)	4.3	0.3	7.3	4.1	7.7
HYDRA-Teacher	54.8	34.1	62.4	55.7	49.4
SFT → GRPO (`Qwen3-8B`, seed 1)	29.0	27.1	36.2	28.3	22.8
ReAct `gpt-5.5`	47.4‡	—	—	—	—

D.3 Agreement with clinician judgment

We compared the scorer's ordering with ratings from one sonographer. The rater reviewed 40 finished episodes: ten per family, with five from each of the top and bottom return quintiles, drawn from eight report-generating policy configurations spanning API-hosted LLM policies, the planner, and learned policies. The rater was blinded to policy identity and score. For each episode, the rater saw the indication, the examination log as the agent saw it, the agent's report and findings, and the source report, and gave 1–5 ratings of acquisition, report, and overall quality.

Ratings agree with the scorer in direction. The report rating correlates with regret at Spearman $\rho = -0.52$ (95% CI $[-0.73, -0.25]$, $n = 40$) and with return at $\rho = +0.45$ $[+0.15, +0.70]$; the overall rating at $\rho = -0.43$ and $+0.33$. Episodes from the top return quintile received a mean overall rating of 3.9 against 2.7 for the bottom quintile. Per family the intervals are wide ($n = 10$); the focused-emergency family, which is synthetic, shows no significant correlation on its own.

The review also audited the scorer. Two of the seven examined disagreements were labeling errors: a clause that only states a size (spleen thickness, breast gland thickness) counted as a positive finding, so a policy that correctly reported a normal spleen was charged a false negative, and one negation pattern ("no flow signal recorded") was read as positive. We corrected both (bare measurements are now judged against reference upper limits) and all reported numbers use the corrected scorer; the change moves family means by at most 2 return units and no ranking. Three disagreements were retryable misses that the scorer prices as licensed false negatives while the rater, seeing a correctly chosen window, did not penalize. The remaining two were episodes from task-trained policies with correct structured verdicts and fragmentary narratives, rated $2/5$: the return scores verdicts, not prose, which is why the report axis is reported separately.

This check uses one rater and a two-tailed sample. It provides no inter-rater estimate, and quintile sampling inflates the rank correlation relative to a uniform sample. The rater also did not clearly separate acquisition from report quality: the two scales correlate at $\rho = 0.87$. The result therefore concerns agreement with an overall clinical impression, not validation of each scoring axis.

E SENSITIVITY TO UTILITY AND SIMULATOR ASSUMPTIONS

The utility magnitudes and observation probabilities are design choices (§ 2.1); the reward construction fixes their ordering, not their clinical calibration. We re-run the four non-LLM policy and reference configurations under ten one-at-a-time perturbations of the illustrative biopsy threshold $p^\star$ that sets $u_{\text{FN}}/u_{\text{FP}}$, the indeterminate utilities, the probe-time weight λ, the plane-acquisition rate and the per-look reveal rate. Belief-based policies plan under the perturbed matrix and the checker scores under it; observation perturbations change the sandbox for every policy. Table 11 reports mean regret in a separate five-seed diagnostic sweep over the evaluation cases. This differs from the aggregation in Table 1. A star marks a change in within-family rank from the default setting; the table reports the associated magnitudes directly.

Table 11: Mean regret under ten single-parameter perturbations. Stars mark rank changes relative to the default setting.

	Thyroid				Breast				Liver				Focused emerg.			
Perturbation	Protocol	Steelman	HYDRA	Bound	Protocol	Steelman	HYDRA	Bound	Protocol	Steelman	HYDRA	Bound	Protocol	Steelman	HYDRA	Bound
default	46.3	25.2	24.6	9.1	32.6	14.7	15.5	3.4	21.2	10.2	2.2	0.9	34.9	32.9	14.9	8.8
$p^\star$=0.02 ($u_{\text{FN}}/u_{\text{FP}}$=49)	82.0	31.6*	32.4*	13.0	59.6	15.7	25.0	4.5	40.7	10.2	3.5	1.2	50.8	33.4	16.8	9.3
$p^\star$=0.10 ($u_{\text{FN}}/u_{\text{FP}}$=9)	34.4	23.0	20.9	7.8	23.6	14.3*	12.4*	3.0	14.7	10.2	1.9	0.8	29.6*	32.8*	12.6	8.6
$u_{\text{IND}} \times 2$	51.7	29.2	28.2	10.3	36.1	17.6*	16.9*	3.9	25.4	12.1	2.3	1.1	45.3	39.0	16.0	10.2
$u_{\text{IND}} \times 0.5$	44.7	23.0	22.8	8.4	31.7	13.1	14.8	3.0	20.7	9.1	2.1	0.8	32.3	29.4	13.2	8.0
λ=0.01	46.6	25.7	25.1	9.5	32.9	15.1	15.5	3.7	21.2	10.4	2.1	1.1	35.1	33.3	14.3	8.8
λ=0.05	45.5	23.6*	24.9*	7.9	31.8	13.6	16.9	2.3	21.2	9.6	3.7	0.2	34.3	31.7	16.8	8.9
p_{plane}=0.8	52.7	25.9	25.1	11.2	36.7	14.8	15.8	4.6	24.7	10.1	2.8	1.8	35.9	32.5	15.6	9.1
p_{plane}=1.0	37.7	23.9	23.9	7.0	23.9	14.6	15.4	2.4	7.3*	10.6*	2.2	0.3	32.2*	34.0*	13.6	8.1
$1-\beta$=0.3	51.5	27.2	26.1	10.5	33.6	15.3	18.1	3.5	20.5	9.5	3.1	0.1	36.2	33.5	19.6	11.8
$1-\beta$=0.7	42.2	23.3	20.7	8.7	31.4	13.4*	11.6*	3.0	21.5	10.6	2.0	1.0	33.7	32.1	12.5	7.5

F EXAMPLE TRAJECTORIES AND REPORTS

We show five thyroid examinations from the evaluation subsample of Table 1. The cases are diagnostic extremes, not a representative sample: two low-regret closed-loop episodes and three high-regret episodes from weaker baselines. Selection uses realized episode regret. For each case, we give the chief complaint, the report of record from which the case was converted (§ 2.2), the recorded trajectory as the environment narrated it, the free-text report the agent filed, its structured verdicts, and the numbers the checker read off the three artifacts. We omit the affordance hints appended to each observation ("may continue: capture image"). The two source-dataset ultrasound images paired with each report are shown for clinical context; they were not exposed to the policy or used as simulator observations. Windows are named as in Appendix B: right lobe, left lobe, isthmus, size/shape, parenchymal echo, capsule/surface, cervical lymph nodes, probe palpation, spectral Doppler. "No usable plane" is the observation model's Ω event; a negative first look at a window that holds a finding is its β event.

F.1 HIGH-SCORING EXAMINATIONS

Case H1. HYDRA (base-rate prior), thyroid case 178763, observation seed 4. Return 28.4 against a bound of 37.2 (regret 8.9); 0 false negatives, 0 unsupported, 0 indeterminate; coverage 100%; QC 78; diagnostic agreement is not applicable because the source impression is empty.

Chief complaint. Neck swelling under investigation.

Report of record. Left lobe $1.6 \times 1.3 \times 1.1$ cm, isthmus 1.7 cm, right lobe $2.0 \times 1.5 \times 1.3$ cm; gland echo slightly heterogeneous; a hypoechoic nodule in the upper-middle right lobe, 2.6×1.5 cm, well defined, regular, with no flow recorded around or within it on CDFI; no enlarged perithyroid lymph nodes.

Source ultrasound images.





Trajectory.

1	B, right lobe	right lobe $2.0 \times 1.5 \times 1.3$ cm
2	B, left lobe	left lobe $1.6 \times 1.3 \times 1.1$ cm
3	B, parenchymal echo	gland echo slightly heterogeneous
4	CDFI, right lobe	*no usable standard plane (right lobe)*
5	CDFI, left lobe	no obvious abnormal flow signal
6	B, size/shape	2.6×1.5 cm; shape regular
7	CDFI, right lobe (re-look)	CDFI: no flow recorded around or within the nodule
8	B, isthmus	isthmus 1.7 cm
9	B, probe palpation	*no usable standard plane (palpation)*
10	B, probe palpation (re-look)	no tenderness on compression
11	B, capsule/surface	capsule intact, surface smooth
12	B, cervical lymph nodes	no enlarged lymph nodes around the thyroid
13	B, spectral Doppler	no abnormal spectral Doppler appearance
14	CDFI, left lobe	no obvious abnormal flow signal
15	capture, size/shape	image stored
16	capture, capsule/surface	*budget exhausted (needs 5, 1 left)*; examination ended: 9 windows (2 CDFI), time 191/192

Report. *Description:* Thyroid: left lobe $1.6 \times 1.3 \times 1.1$ cm. Isthmus 1.7 cm. Right lobe $2.0 \times 1.5 \times 1.3$ cm. Gland echo slightly heterogeneous. Size 2.6×1.5 cm. Shape regular. CDFI: no flow recorded around or within the nodule. No enlarged lymph nodes around the thyroid. No obvious abnormal flow signal. Capsule intact, surface smooth. *Impression:* Thyroid and neck ultrasound: parenchymal echo change, diffuse disease possible; suggest correlation with laboratory tests and follow-up.

Structured verdicts. Present: parenchymal echo, isthmus, left lobe, right lobe, size/shape, right-lobe flow. Absent: capsule, lymph nodes, left-lobe flow. Indeterminate: none.

Reading. The first Doppler look at the right lobe returned no plane (step 4). The agent moved on, the size/shape window revealed the nodule (step 6), and the posterior sent it back to the right lobe for the Doppler read it now valued (step 7). Every question is answered from evidence. The prose is a concatenation of revealed fragments: the nodule's size floats without its location, and the impression is a template sentence about diffuse disease that neither monitor penalizes. The narrative is the weakest of the three artifacts in this case, consistent with § 5.

Case H2. Closed loop without belief (re-look + abstain), thyroid case 182695, observation seed 3. Return 25.3 **against a bound of** 29.6 **(regret** 4.3**); 0 false negatives,** 0 **unsupported,** 2 **indeterminate; coverage** 88% **; QC** 77**, diagnostic agreement not applicable because the ground-truth impression is empty.**

Chief complaint. Neck swelling under investigation.

Report of record. Thyroid slightly small in volume, gland echo heterogeneous, no abnormal flow within the gland on CDFI; no enlarged cervical lymph nodes bilaterally.

Source ultrasound images.




Trajectory.

1	B, size/shape	thyroid volume slightly small
2	B, capsule/surface	capsule intact, surface smooth
3	B, parenchymal echo	gland echo heterogeneous
4	B, left lobe	no obvious abnormality
5	B, right lobe	no obvious abnormality
6	B, isthmus	isthmus not thickened
7	B, cervical lymph nodes	no enlarged lymph nodes bilaterally
8	B, capsule/surface	*no usable standard plane (capsule)*
9–10	B, capsule/surface ×2	capsule intact, surface smooth (both)
11–12	B, left lobe ×2	no obvious abnormality (both)
13–14	B, right lobe ×2	no obvious abnormality (both)
15–16	B, isthmus ×2	isthmus not thickened (both)
17	capture, size/shape	image stored
18	capture, capsule/surface	*budget exhausted (needs 5, 3 left)*; examination ended: 7 windows (0 CDFI), time 189/192

Report. *Description:* Thyroid volume slightly small. Gland echo heterogeneous. No enlarged cervical lymph nodes bilaterally. *Impression:* (none filed).

Structured verdicts. Present: size/shape, parenchymal echo. Absent: capsule, isthmus, left lobe, right lobe. Indeterminate: lymph nodes, and the flow question, which was never examined.

Reading. A benign case with no window decision to make, so the belief-free planner is close to the reference. Its evidence floor spends the second half of the budget on repeat looks at windows that had already answered (steps 9–16) and never switches to CDFI, which is the one question it leaves open and the principal source of its regret. It declines the lymph-node question in the structured verdict after a single look, while the prose assembled from the same fragments asserts it. This is the disagreement measured by *Contr.* in Table 2.

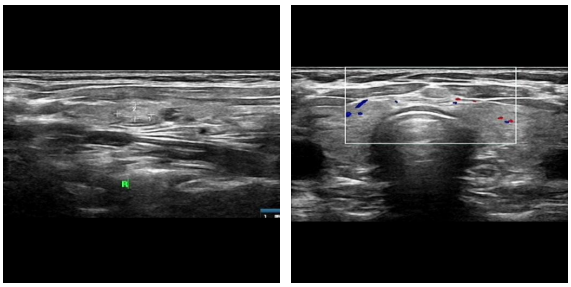
F.2 Low-scoring examinations

Case L1. Open-loop fixed protocol, thyroid case 176508, observation seed 2. Return -57.7 against a bound of 37.2 (regret 95.0); 3 false negatives, 2 unsupported, 1 indeterminate; coverage 62%; QC 44; diagnostic agreement is not applicable.

Chief complaint. Thyroid nodule found on a routine check-up; referred for further work-up.

Report of record. Thyroid normal in size and shape; a slightly hyperechoic nodule in the upper right lobe, 1.6×1.4 cm, and a slightly hypoechoic nodule in the isthmus, 1.6×1.3 cm, both well defined and regular with no flow on CDFI; the remaining gland homogeneous with no abnormal flow; no enlarged cervical lymph nodes.

Source ultrasound images.



Trajectory.

1	B, size/shape	*no usable standard plane (size/shape)*	
2	B, capsule/surface	*no usable standard plane (capsule)*	
3	B, parenchymal echo	remaining gland echo homogeneous	
4	B, left lobe	no obvious abnormality	
5	B, right lobe	no obvious abnormality	(*nodule not shown on this look*)
6	B, isthmus	isthmus not thickened	(*nodule not shown on this look*)
7	B, cervical lymph nodes	no enlarged lymph nodes bilaterally	
8–11	capture, size/shape, capsule, echo, left lobe	images stored	
12–13	capture, right lobe $\times 2$	*no usable plane; capture failed* (both)	
14–16	capture, right lobe, isthmus, lymph nodes	images stored; examination ended: 7 windows (0 CDFI), time 119/192	

Report. *Description:* Remaining gland echo homogeneous. No enlarged cervical lymph nodes bilaterally. *Impression:* (none filed).

Structured verdicts. Absent: size/shape, capsule, parenchymal echo, left lobe, right lobe, isthmus, lymph nodes. Present: none. Indeterminate: the flow question.

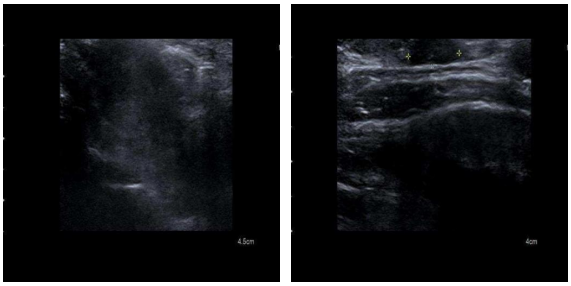
Reading. Two windows never yielded a plane (steps 1–2) and the report declares both absent regardless: the two unsupported negatives, made through a closed window (the *no usable plane* column of Table 8). The right-lobe and isthmus nodules were missed on a single look each (β, steps 5–6) and never revisited, although 38% of the budget was left unspent. The complaint named a nodule; the examination found none. This is the failure mode excluded by Proposition 1's reliable, concentrated regime: a fixed list encounters closed windows and then asserts through them.

Case L2. Zero-shot policy (`Qwen3-8B`), thyroid case 195274, observation seed 1. Return -15.3 against a bound of 41.0 (regret 56.3); 0 false negatives, 1 unsupported positive, 0 unsupported negatives, and 8 indeterminate; coverage 0%; QC 42; diagnostic agreement is not applicable.

Chief complaint. Palpable left neck lump for two weeks.

Report of record. Thyroid normal in size and shape, left lobe 1.8 and right lobe 2.0 cm thick, isthmus about 1.6 cm; gland echo heterogeneous with multiple patchy slightly hypoechoic areas and slightly increased flow on CDFI; a hypoechoic nodule below the left lower pole, 2.0×1.7 cm, well defined, regular, hilum not clearly shown, no flow; a hypoechoic mass below the right lower pole, 1.8×1.2 cm, fairly well defined, fairly regular, with flow on CDFI.

Source ultrasound images.



Trajectory.

1–12	B, left lobe $\times 12$	"lymphatic hilum" not clearly shown (the same fragment, twelve times)
13	B, left lobe	*budget exhausted (needs 16, 0 left)*; examination ended: 1 window (0 CDFI), time 192/192

Report. *Description:* A mass is seen in the left lobe, ill defined, irregular, with heterogeneous internal echo, no clear capsule and no obvious flow signal. The mass is mobile on probe palpation. Cervical lymph nodes unremarkable. *Impression:* Left-lobe mass, most likely benign; suggest further work-up or follow-up.

Raw structured output (eight finding objects). Present: the left-lobe mass and its mobility. Absent: left-lobe flow. Indeterminate: left-lobe margins, shape, parenchymal echo and capsule, and the cervical lymph nodes. These are the model's submitted objects, not yet the scored question vector.

Mapping to scored questions. The scorer maps the objects to nine canonical questions. The left-lobe objects collapse onto the B-mode left-lobe question; because the submitted flow object has `is_flow=false`, it does not populate the separate left-flow question. The resulting vector is PRESENT for the left-lobe question and INDETERMINATE for the other eight questions. Since all twelve looks returned `plane_ok=false` with no observations, the left-lobe positive is unsupported. Thus the final tally is one unsupported positive, zero unsupported negatives, and eight indeterminate verdicts.

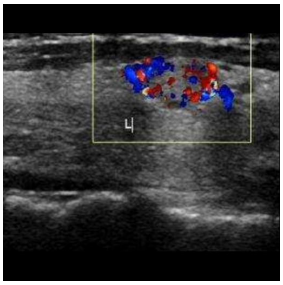
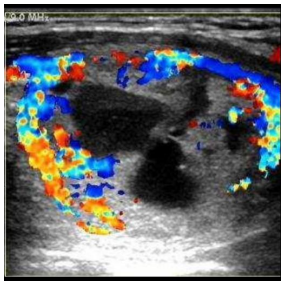
Reading. The zero-shot policy repeats one action until the clock runs out: twelve looks at the left lobe, none of which returns a usable plane, and a thirteenth refused for budget. Coverage is therefore 0%. The report is nevertheless unfaithful: the unsupported positive left-lobe mass follows the chief complaint rather than acquired evidence, and the prose adds morphology and mobility that the trajectory never revealed. Most of the regret comes from failing to acquire evidence and leaving questions open. The right-lobe mass and diffuse change are never sought. QC 42 mainly reflects missing images and an incomplete report; that monitor does not encode the one-window trajectory.

Case L3. Zero-shot LLM-based policy (ReAct, `gpt-4o-mini`), thyroid case 184525, observation seed 5. Return -49.8 against a bound of 62.4 (regret 112.2); 2 false negatives, 0 unsupported, 6 indeterminate; coverage 33%; QC 34; diagnostic agreement is not applicable.

Chief complaint. Palpable right neck lump for six months.

Report of record. Left lobe normal in size and shape, right lobe enlarged, gland echo heterogeneous; a cystic-solid mass in the right lobe, $1.4 \times 0.9 \times 0.7$ cm, well defined, regular, with a hypoechoic halo and rich flow around and within it on CDFI; a hypoechoic nodule in the mid left lobe, 0.6×0.5 cm, well defined, regular, with rich flow throughout; no enlarged cervical lymph nodes.

Source ultrasound images.

Trajectory.

1	B, right lobe	no obvious abnormality	(*mass not shown on this look*)
2	B, left lobe	no obvious abnormality	(*nodule not shown on this look*)
3	B, isthmus	isthmus not thickened	
4	finish	examination ended: 3 windows (0 CDFI), time 40/192	

Report. *Description:* Thyroid ultrasound shows no obvious abnormal echo in the right lobe, left lobe or isthmus; no mass or thickening. *Impression:* No obvious thyroid lesion on ultrasound; suggest regular follow-up.

Structured verdicts. Absent: right lobe, left lobe, isthmus. Indeterminate: lymph nodes, and the five questions it never examined.

Reading. The complaint names the right side. The agent looks once at each lobe, accepts a negative first look on the right (β, step 1), never switches to CDFI, and finishes with 79% of the budget unspent. Its three negatives are licensed, so nothing is unsupported and the report axis is clean; both lesions are missed and six questions are left open. It fails by not looking, which is the pattern of the ReAct row in Table 1.