REINFORCEMENT LEARNING IN PARTIALLY VERIFIABLE ENVIRONMENTS FOR COLON CANCER TREATMENT PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical decision-making is a sequential process of information collection, clinical inference, and forecasting under partial observability, often without a uniquely correct treatment decision. Large language models (LLMs) have demonstrated strong medical question-answering capabilities but extending these advances to clinical reasoning remains challenging because many clinical tasks are only partially verifiable. Recent clinical reinforcement learning (RL) work has therefore largely focused on diagnostic endpoints, where correctness can be more directly verified. Oncology treatment decision making is a particularly challenging instance of this problem that extends beyond medical question-answering in which multiple decision trajectories can be clinically defensible for the same partially observed patient state. We formulate oncology treatment planning as an interactive RL problem under partial observability and construct patient-specific, process-aware rewards from decision processes observed in real world clinical records. The agent must learn to sequentially acquire information from a structured patient representation before committing to a first-line treatment recommendation, and is evaluated on information gathering, treatment selection, and reasoning alignment. Using GRPO and GDPO to optimize Llama-3.1-8B-Instruct within this environment, RL substantially improves the base policy beyond frontier model capabilities. The GDPO-trained policy increases Mean@3 scalar clinical reward by 79.60% over base Llama and exceeds GPT-5.2 by 18.82% under the same interactive environment, while reducing query count during training. Overall, our results show that patient-specific, partially verifiable rewards can provide an effective learning signal for interactive clinical RL when both the patient state and the underlying clinical decision process are only partially observed [1].

1 INTRODUCTION

Clinical artificial intelligence has advanced rapidly in its capabilities across clinical reasoning tasks such as medical question-answering (Jin et al., 2019; 2021; Pal et al., 2022), clinical inference (Tu et al., 2025; Chen et al., 2024; 2026), and clinical forecasting (Ma et al., 2024). Existing benchmarks primarily focus on medical question-answering and treatment planning, and often present the model with a curated clinical scenario in which the information relevant to the decision has already been supplied (Kanjee et al., 2023).

Yet many of the most consequential problems in medicine extend beyond identifying a diagnosis to deciding what should be done next when information is only partially observed (Arora et al., 2025; Chiu et al., 2026; Nori et al., 2025; Li et al., 2024; Vasilev et al., 2025; Schmidgall et al., 2026; Jiang et al., 2025). Clinical decision-making requires physicians to determine what clinical information is necessary, integrate evidence acquired from multiple sources, and select among interventions whose benefits and risks depend on the particular patient. AI systems intended to support these decisions must therefore do more than map a completed case description to a recommendation; they must participate in the clinical reasoning process through which the information required for a decision is acquired, interpreted and translated into action (Qiu et al., 2025b; Rao et al., 2026).

[1] Anonymized codebase can be found here.

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.

Even after the relevant evidence has been acquired, there may be no uniquely correct treatment action, with different clinicians reasonably coming to different treatment choices. Multiple prospective treatment plans may remain clinically defensible because treatment decisions reflect tradeoffs among efficacy, toxicity, contraindications, treatment sequencing, patient preferences, feasibility, and other contextual constraints such as institutional regulations and compliance (Krell et al., 2015). Thus, evaluating treatment-planning quality requires assessing not only the final recommendation, but also whether the clinical agent acquired and used the information necessary to justify that recommendation (Ruhrberg Estévez et al., 2026).

The best data available for oncology treatment research is based on real decisions and outcomes within longitudinal clinical records, but even these do not necessarily provide ground-truth annotations of the underlying decision process. In particular, they do not specify which information should be acquired at each intermediate state or which clinical inference should follow from the accumulated evidence. Even at the point of treatment selection, the historically observed conditions need not constitute a unique ground-truth action; different clinician phenotypes can give rise multiple clinically defensible first-line treatment plans for the same patient. In this work, we therefore focus on immediate clinical decision-making: an agent sequentially acquires patient-specific information, performs clinical inference over the acquired evidence, and ultimately selects a first-line treatment. The resulting supervision is therefore naturally expressed as reward over a sequence of decision.

Our central question is: *how can we design strongly verifiable reinforcement-learning environments when the clinically relevant patient state is partially observed and the available clinical signals provide only partially verifiable rewards?*

Here, we use environment to denote the complete interactive system through which a clinical agent acquires and infers information about the underlying patient state and receives rewards. Most importantly for the setting considered here, environment design determines what aspects of an agent's trajectory can be verified and how those verification signals are translated into rewards. Rather than treating verifiability as binary, we consider rewards along a continuum from weakly to strongly verifiable, according to how unambiguously and reliably the environment can determine whether an agent's behavior satisfies the relevant clinical criteria.

Our main contributions are threefold.

1. We formulate oncology treatment planning as an interactive reinforcement-learning problem under partial observability and construct an environment from retrospective clinical records in which an agent must actively acquire patient-specific information before committing to a treatment recommendation. We couple this interaction with a patient-specific, process-aware reward that evaluates information gathering, treatment selection, and alignment with the documented clinical rationale.

2. We show that reinforcement learning within this environment can produce a relatively small, specialized language model that substantially outperforms a frontier general-purpose model on the same interactive oncology treatment-planning task. Under the same patient environment, retrieval interface, turn budget, and evaluation procedure, the RL-trained Llama-3.1-8B policy achieves higher treatment-planning reward than GPT-5.2. These results suggest that appropriately constructed domain-specific environments can provide learning signals strong enough for smaller models to acquire specialized sequential decision-making behavior not exhibited by substantially larger models zero-shot.

3. We provide evidence that the learned policy acquires clinically structured decision-making behavior rather than improving only its final-answer accuracy. During reinforcement learning, the policy improves its overall reward while reducing the number of queries it issues, consistent with learning to acquire more decision-relevant information. Its information-seeking trajectories, treatment rationales, and recommendations are also concordant with the documented decision processes of oncologists. In a preliminary clinician comparison, the specialized policy performs comparably to an oncologist under the same interaction protocol; because this comparison currently contains only a small number of cases, we treat it as suggestive evidence rather than evidence of clinician-level performance.

2 Related Work

Interactive clinical agents and sequential clinical decision-making. Recent work has extended clinical language models from single-response reasoning to agents that interact sequentially with clinical environments. Similarly, recent clinical RL methods optimize information acquisition. Doctor-R1 and DoctorAgent-RL train agents through multi-turn patient dialogue (Lai et al., 2026; Feng et al., 2026). DiagAgent, SALUS and LA-CDM optimize the policy that seeks information sequentially until it is sufficient confident with the diagnosis (Qiu et al., 2025a; Gao et al., 2026; Bani-Harouni et al., 2026). ResidencyRL broadens this paradigm to long-horizon simulated encounters involving inquiry, tool use, documentation, diagnosis, management, and safety (Liévin et al., 2026). Most of these environments remain organized around diagnostic endpoints. In contrast, our environment places the patient-specific clinical state itself under partial observability and focuses on improving the agent's reasoning capabilities required to sequentially acquire and interpret evidence from an initially hidden patient record before committing to a first-line treatment decision.

Reinforcement learning under partially verifiable clinical feedback. Much of the recent reinforcement learning (RL) for reasoning assumes setting in which either final outcome is verifiable or intermediate reasoning steps can be explicitly supervised, such as mathematics and programming Lightman et al. (2024); Jia et al. (2025). A complementary line of work studies how RL can be extended beyond tasks with deterministic verifiable signals, especially in the clinical setting. Clinical-R1 introduces multi-objective clinical RL that jointly rewards correctness, faithfulness, and comprehensiveness (Gu et al., 2026), while Rubrics as Rewards replaces binary verification with instance-specific, multi-objective feedback based on natural language for on-policy RL (Gunjal et al., 2026). More recently, GatorOnco applies RL to colorectal-cancer treatment planning, using a hierarchical treatment reward together with clinical-guideline retrieval to promote guideline-concordant treatment generation (Lyu et al., 2026). On the other hand, our reward is anchored to patient-specific decision processes reconstructed from retrospective clinician notes rather than primarily to guideline concordance. Because neither intermediate reasoning actions nor the terminal treatment necessarily admit a unique ground-truth label, we use a terminal but process-aware reward that evaluates the complete interactive trajectory across information acquisition, treatment selection, and clinical reasoning.

3 Clinical Problem Formulation

First-line treatment planning for patients with colon cancer requires oncologists to integrate disease characteristics, molecular biomarkers, prior treatment history, patient functional status, comorbidities, preference, and treatment intent to determine an appropriate initial management strategy. In clinical practice, decisions are made under incomplete information and unfold through iterative and intertwined phases of data acquisition and clinical inference. Oncologists must identify and acquire the evidence needed to distinguish among potentially defensible treatment options, while continuously interpreting newly acquired evidence to refine subsequent information gathering and candidate treatment plans. These processes are sequential but tightly coupled, iterating until sufficient evidence has been gathered to support a treatment decision. We therefore formulate this process as a stochastic sequential decision-making problem under partial observability (see Figure 3).

We reward the agent against the documented per-patient decision trajectory, rather than explicit adherence to published clinical guidelines. We do this in recognition that these guidelines are just that: a starting guide for treatment, but often deviated from actual clinical practice. These deviations due to the nuanced nature of clinical reality are of interest in training models that can replicate the realities of clinical decision making. In addition, the extracted decision trajectory is evidence of one clinically realized decision. There is no unique trace of information-gathering and reasoning process with stepwise label, and for the same patient state, multiple clinically reasonable treatment options can be available. We therefore formulate clinical verifiability as inherently partial: some properties of a trajectory can be evaluated against patient-specific evidence and the documented oncologist decision process, while the overall quality of the trajectory cannot be reduced to agreement with a single uniquely correct sequence of actions.

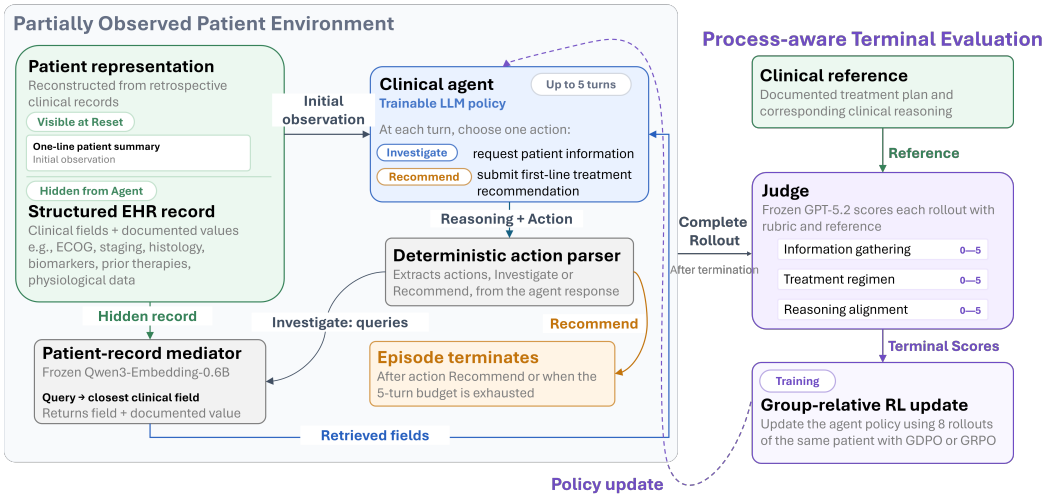


Figure 1: The system consists of five functional components: a patient environment constructed from retrospective clinical data, a trainable clinical agent, a deterministic action parser, a frozen patient-record mediator, and a frozen judge.

4 Methodology for Clinical Reinforcement Learning Environment Design

Here, we devise an interactive RL environment for colon cancer treatment planning under partial observability and partially verifiable clinical feedback. While what we describe herein is applied to oncology information gathering and decision making, the system is generalizable to RL construct across the broader medical domain. The system is comprised of a trainable clinical agent interaction with a patient representation through an deterministic action parser and frozen patient-record mediator, while a frozen judge evaluates the completed trajectory. At the start of the trajectory, the agent will receive a one-line summary regarding the patient. The agent provides reasoning and actions between investigation, which is passed through the parser and patient-record mediator to retrieve patient-specific clinical information and return it to the agent, or a terminal treatment recommendation. The completed trajectory is evaluated by the frozen judge using a process-aware reward model, with the resulting terminal reward converted into a group-relative advantage for optimizing the clinical agent's policy. We provide an overview of the system in Figure 1. We describe each component in detail below, but defer the experimental setup and prompts to Appendices E and J.

4.1 Data & Cohort Construction

We construct an oncology treatment-planning dataset from retrospective, longitudinal, de-identified electronic health records (EHR)s from a large academic medical center in the United States, comprising 950 adult colon cancer patients. We extracted documented first-line treatment plan and their underlying clinical rationales, including the information acquisition process of the oncologists, to supervise our process-aware reward model. To build an interactive environment, we reconstructed longitudinal patient trajectories into a structured representation of history of present illness, oncologic profiles, and physiological data (Chiu et al., 2026). The dataset is split at the patient level into disjoint training, validation, and test sets, comprising 665, 143, and 142 patients. Full details on cohort inclusion and exclusion criteria, patient representation schema definition, and extraction prompts can be found in Appendix A.

4.2 Patient Representation and Environment State

We index patients by i, trajectory by g, and interaction turns by $k \in \{0, \ldots, K-1\}$, where K denotes the turn budget. Let χ^i denote the underlying clinical state of the patient i and $\rho^i = \rho\left(\ell^i, \mathcal{R}^i\right)$ where ℓ^i is a one-line patient summary available at the start of the episode and $\mathcal{R}^i$ is the patient-

specific EHR record which is initially hidden but available through the retrieval. Because ρ^i is constructed from documented clinical data, it is an incomplete proxy for the underlying clinical state χ^i.

No clinical time elapses during a trajectory. The agent's actions do not change the patient's physiology or generate counterfactual treatment outcomes. As a result, the environment dynamics arise from the sequential information acquisition, clinical inference and decision planning.

To facilitate rapid RL training and minimize cost, we converted the unstructured clinical not documenting patient state and clinician decision to a structured key-value dictionary. The retrieval record is represented as $\mathcal{R}^i = \left\{\left(\xi, v^i(\xi)\right) : \xi \in \mathcal{F}^i\right\}$, where $\mathcal{F}^i$ denote clinical fields available for patient i, ξ is a clinical field and $v^i(\xi)$ is its documented value.

The formulation of the partially observed Markov decision process will require us to know the fully observed RL trajectory $\omega^{i,g}$ which can be constructed from the fully observed state-action history $H_k^{i,g} = (s_0, a_0, s_0, a_1, ..., s_k)$, the patient representation ρ^i, the budget K and any additional information flow x that passes through the environment (parser and mediator) which the agent does not observe. The fully observed RL trajectory is then parametrized by

$$\omega_k^{i,g} = \left(H_k^{i,g}, \rho^i, K - k, x\right). \tag{1}$$

We leave x as an placeholder symbol for the rest of the environment information that goes to the judge model but the agent does not observe. We further remark that we will formulate the reward model as a terminal reward that does not appear within the history $H_k^{i,g}$ until the terminal step. Full state and trajectory notation is provided in Appendix B.

4.3 Clinical Agent Action Space, Structured Action Parser & Frozen Patient-Record Mediator

At the turn k, the clinical agent with a parameterized policy π_{θ_n} conditioned on the observed transcript generates a natural-language response $a_k^{i,g}$, which denotes the complete token sequence generated at one turn.

We equip the clinical agent with a deterministic parser that maps each response to either an investigation action or a terminal treatment recommendation. Investigation actions may contain multiple queries, which are screened by deterministic regular expressions to reject requests for clinical inference and retain only requests for clinical variables. Accepted queries are executed independently, while corrective messages are returned for rejected queries. If no valid structured action can be extracted, then no clinical action is executed and the turn is consumed. In order to design the environment in such a way that it is auditable, we equip the agent with a retrieval harness that keeps track of the documents it retrieved. The patient-record mediator maps any accepted action request to the most semantically similar field in the patient representation and returns the associated documented value. We detail the specific of the parser and the mediator in Appendix C.

We note that the the mediator performs a top-1 retrieval without any abstention threshold. Every accepted request therefore receives a key-value response, even if the intended clinical variable was absent from $\mathcal{F}^i$. In that case, the mediator returns the nearest available field rather than indicating the requested variable was missing. This behavior does introduce semantic retrieval error into the observation channel but it is conditioned on a fixed request, patient representation and the frozen mediator and hence the mapping is deterministic and it is just an approximation error rather than a stochastic observation noise. This error, however can be traced back to its origin because one may always check which clinical field was actually retrieved so auditability is not lost.

4.4 Reward Model Construction

Treatment planning does not admit an exact verifier like in the case of programming or mathematics. Although the treatment selected by an oncologist in our dataset provides a signal of a clinically realized decision, it is not necessarily true that it is the only clinically defensible treatment.

We therefore evaluate each trajectory against a patient-specific clinician reference ζ_{ref}^i. The clinician reference contains the documented treatment plan and the corresponding clinical reasoning. The

reference is withheld from the clinical agent throughout the interaction with the environment and is supplied only to the judge after termination. We do not provide the agent a reward at each non-terminal turn. Hence, the reward model gives terminal rewards but remains process-aware.

The judge receives $\left(\omega^{i,g}, \zeta^{i}_{\text{ref}}, \mathcal{U}\right)$, where $\mathcal{U}$ is a fixed evaluation rubric meant to evaluate the trajectory. The judge then produces three bounded integer scores

$$r^{i,g}_{\text{info}}, \quad r^{i,g}_{\text{regimen}} \quad r^{i,g}_{\text{reason}} \in \{0, 1, \ldots, 5\}. \tag{2}$$

The information-gathering score $r^{i,g}_{\text{info}}$ evaluates whether the clinical agent acquired the information needed for the treatment decision. The score penalizes both premature recommendation and inefficient over-investigation.

The regimen-accuracy score $r^{i,g}_{\text{regimen}}$ evaluates the terminal treatment recommendation against the clinician reference. Exact agreement receives the highest score, while clinically equivalent alternatives may receive substantial credit. Selecting the wrong treatment class or an incompatible regimen receives substantially less credit than omitting a secondary component.

The reasoning-alignment score $r^{i,g}_{\text{reason}}$ evaluates whether the agent correctly connects the acquired patient-specific evidence to its recommendation. It measures whether the agent uses relevant variables consistently, interprets them correctly, and appropriately includes or excludes treatment alternatives. This score distinguishes a patient-specific recommendation from an otherwise similar recommendation unsupported by the evidence acquired during the episode.

We define a scalar clinical reward as

$$r^{i,g}_{\text{clin}} = \frac{r^{i,g}_{\text{info}} + r^{i,g}_{\text{regimen}} + r^{i,g}_{\text{reason}}}{15} - 0.2 \cdot \mathbb{1}_{\text{p}} - 0.4 \cdot \mathbb{1}_{\text{c}}\,, \tag{3}$$

where $\mathbb{1}_p$ is an indicator that is nonzero if there is any error message from the parser. 1_c is a completeness indicator that is nonzero if the agent did not output a treatment recommendation before the turns run out. Each indicator is a penalty that can only be activated once per trajectory.

4.5 Objectives and Algorithms: GRPO and GDPO

Group Relative Policy Optimization (GRPO) is suitable for our setting because each patient-specific environment can generate multiple trajectories under the same underlying patient representation (Shao et al., 2024). These trajectories form a rollout group, allowing the relative quality of different information-gathering, reasoning, and treatment trajectories for the same patient to be estimated without learning a separate value function.

Since GRPO first scalarizes the reward components into one terminal reward, then performs group normalization, and finally calculates the relative advantage for each trajectory within the rollout group, it cannot distinguish two rollouts with different clinical reward profiles but the same scalar total. For example, strong information gathering followed by an incorrect regimen may receive the same scalar reward as weak information gathering followed by an accurate regimen. Group reward-Decoupled Normalization Policy Optimization (GDPO) addresses this loss of reward resolution by normalizing each clinical reward component separately before combining the resulting component advantage (Liu et al., 2026). Our hypothesis is that GDPO will be better able to learn the combined tasks of information gathering and decision making than GRPO due to its ability to distinguish between the reward components during advantage calculation.

For both GRPO and GDPO, we sample G trajectories for each patient and construct the policy update from their group relative advantage derived from the terminal rewards. The same trajectory-level advantage is broadcast to every agent-generated token in that rollout. Only tokens generated by the clinical agent contribute to the policy-gradient term. We mask the tokens belonging to the system prompt, initial patient observation, parser messages, or mapper response. A full description on the advantage formulation and the objective function is provided in Appendix D.

5 Results

We evaluate on 142 held-out patients with three generations per case. We report the Mean@3 and Best@3 of the reward calculated using LLM-as-a-judge as described in Section 4.4. Mean@3 de-

Table 1: Overall reward performance across different agent policies. Results are evaluated on $n = 142$ cases with three generations per case. Mean@3 is the mean scalar reward across three generations; Best@3 is the mean of the highest scalar reward among the three generations for each case. Best results are shown in **bold** and second-best results are underlined.

	Mean@3	Best@3	SD	95% CI
GRPO Llama-3.1-8B-Instruct	<u>0.5712</u>	0.6446	0.1465	[0.547, 0.595]
GDPO Llama-3.1-8B-Instruct	**0.6085**	**0.6718**	0.1264	[0.588, 0.629]
GRPO Qwen3.5-9B	0.1994	0.4878	0.2096	[0.165, 0.234]
GDPO Qwen3.5-9B	0.3429	0.5808	0.2174	[0.307, 0.379]
GRPO GLM-4-9B-0414	0.5313	0.6263	0.1321	[0.510, 0.553]
GDPO GLM-4-9B-0414	0.5105	0.6192	0.1474	[0.486, 0.535]
GPT-5.2	0.5121	<u>0.6469</u>	0.2408	[0.472, 0.552]
Llama-3.1-8B-Instruct	0.3388	0.4934	0.1422	[0.315, 0.362]
Qwen3.5-9B	−0.0142	0.1474	0.1505	[-0.039, 0.011]
GLM-4-9B-0414	−0.0163	0.1981	0.1879	[-0.047, 0.015]

notes the average scalar clinical reward across the three generations per case, while Best@3 averages the highest scalar clinical reward among the three generations per case.

RL improves all three policy backbones, although the magnitude of improvement varies substantially by backbone. For Llama-3.1-8B-Instruct, GRPO and GDPO increase Mean@3 reward from 0.3388 to 0.5712 and 0.6085, respectively. Qwen3.5-9B improves from -0.0142 to 0.1994 with GRPO and 0.3429 with GDPO, while GLM-4-9B-0414 improve from -0.0163 to 0.5313 with GRPO and 0.5105 with GDPO. Among the evaluated policies, GDPO-trained Llama achieves the highest Mean@3 reward and exceeds GPT-5.2, which attains 0.5121 (see Table 1).

Reward decomposition shows that RL improves Llama across information gathering, treatment selection, and reasoning relative to its base policy, while its relative advantage over GPT-5.2 is concentrated in treatment selection. GRPO-trained and GDPO-trained Llama achieve treatment rewards of 3.366 and 2.791, respectively, compared with 2.702 for GPT-5.2. In contrast, GPT-5.2 remains stronger on information gathering and reasoning with Mean@3 total scalar rewards of 2.977 and 3.016, respectively. GRPO-trained Llama achieves 2.754 and 2.688 on these axes, while GDPO-trained Llama achieves 2.953 and 2.791 (see Tables 2 and 4).

Across Llama training, treatment reward increases more strongly than information-gathering and reasoning reward, while GRPO exhibits a reward collapse at the later stage of the training as shown in Figure 2. Information-acquisition behavior also changes with RL, although not uniformly across all the policy backbone (see Figures 5 and 6).

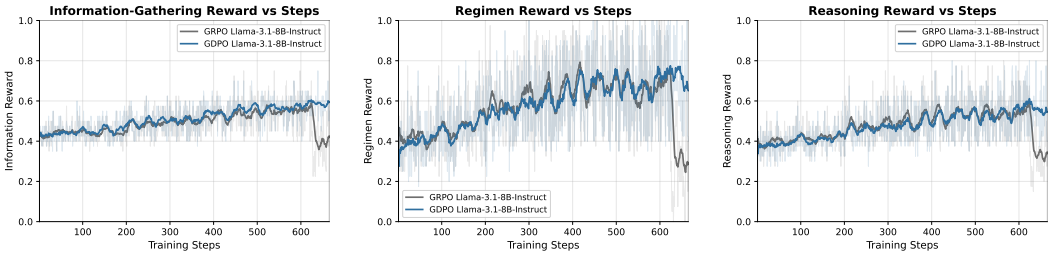
To compare the performance of the agent with human, we recruit four oncologists to produce trajectories in the same interactive environment. Each clinician is assigned with 6 to 10 unique cases from the held-out test set. We allow the clinician to send out probe to the patient representation and leave the reasoning in the same fashion as the agent does. Clinicians need to provide a treatment recommendation within 5 turns. We report the average reward across all axes compared to the models on the same patient set. On the same 34 held-out cases performed by four oncologists, GDPO-trained Llama attains an average scalar reward of 0.6000, compared with 0.573 for the oncologists, 0.544 for GPT-5.2, and 0.537 for GRPO-trained Llama (see Table 3).

6 Discussion & Conclusion

Clinical treatment planning requires decision-making under partial observability, yet existing clinical language model benchmarks and reinforcement learning environments often assume complete information, evaluate fixed policies, or focus primarily on diagnostic endpoints. In this work, we formulate oncology treatment planning as an interactive reinforcement learning problem and construct an environment from retrospective clinical records in which an agent must acquire patient-specific information before making a treatment recommendation, with a process-aware reward evaluating information gathering, treatment selection, and clinical reasoning. Reinforcement learning withing

Table 2: Mean@3 performance on each reward axis. Information, treatment, and reasoning are scored from 0–5; completeness and format are scored from 0–1. Best results are shown in **bold** and second-best results are underlined.

Policy	Information (0–5)	Regimen (0–5)	Reasoning (0–5)	Complete (0–1)	Format (0–1)
GRPO Llama-3.1-8B-Instruct	2.754	3.366	2.688	0.960	1.000
GDPO Llama-3.1-8B-Instruct	2.953	**3.453**	2.791	**0.988**	1.000
GRPO Qwen3.5-9B	2.606	1.453	2.298	0.439	1.000
GDPO Qwen3.5-9B	2.725	2.033	2.399	0.664	1.000
GRPO GLM-4-9B-0414	2.728	2.967	2.444	0.972	1.000
GDPO GLM-4-9B-0414	2.629	2.854	2.455	0.953	1.000
GPT-5.2	**2.977**	2.702	**3.016**	0.831	1.000
Llama-3.1-8B-Instruct	2.124	1.876	1.927	0.859	1.000
Qwen3.5-9B	2.406	0.437	2.085	0.143	1.000
GLM-4-9B-0414	2.002	0.678	1.723	0.225	1.000



(a) Information-gathering reward. (b) Treatment reward. (c) Reasoning reward.

Figure 2: Learning dynamics of GRPO and GDPO with Llama-3.1-8B-Instruct. Dark curves show a 15-step moving average and faint curves show the unsmoothed training trajectory. Each curve corresponds to a single training run.

this environment produces a specialized Llama-3.1-8B policy that attains higher overall treatment-planning reward than both its base policy and GPT-5.2, while improving across all three clinical reward axes and learning to acquire information more selectively. Overall, these results demonstrate that retrospective clinical data can support interactive reinforcement-learning environments that provide sufficient learning signal to improve sequential clinical decision-making, even when both the patient state and the underlying clinical decision process are only partially observed and the reward is only partially verifiable.

Formulating treatment planning as an interactive problem allows us to decompose the expert reasoning process into several distinct phases with observables. In the information acquisition phase, the policy determines which information to seek from the partially observed environment. The observable behaviors then include the number of queries and the clinical fields retrieved, and the sequence in which information is acquired. Subsequently, the policy uses the retrieved evidence to determine whether to perform further investigation or whether sufficient information is available to make a treatment recommendation. We refer to this iterative process of information acquisition, inference, and forecasting as expert reasoning. The decomposition of rewards into information gathering, treatment, and reasoning in our environment allows us to observe these distinctions between policies.

The training dynamics further suggest that reinforcement learning changes the underlying clinical policy rather than only improving the terminal treatment recommendation. The GRPO-trained Llama policy improves its clinical reward while reducing the number of queries it issues, consistent with learning a more selective information-acquisition policy rather than merely increasing the volume of retrieved information (see Table 5). At the same time, the different reward axes improve at different rates, with treatment reward increasing more rapidly than information-gathering and reasoning rewards. Together, these observations suggest that interactive clinical RL environments make it possible to study how a base policy gradually adapts toward expert reasoning during training.

Table 3: Comparison between oncologists and language-model policies in the same interactive treatment-planning environment. Four oncologists each evaluated 6 to 10 unique cases from the held-out test set, for a total of 34 clinician trajectories. Agent results are evaluated on the same 34 cases under the same five-turn interaction protocol and averaged over three generations per case. Information, treatment, and reasoning rewards are scored from 0–5; completeness is scored from 0–1.

Policy	n	**Scalar Reward**	**Information** (0–5)	**Regimen** (0–5)	**Reasoning** (0–5)	**Complete** (0–1)	**Queries per case**
Oncologists	34	0.573	3.000	2.912	2.676	1.000	9.412
GRPO Llama	34	0.537	2.647	3.206	2.608	0.931	4.598
GDPO Llama	34	0.600	2.931	3.382	2.804	0.980	11.275
GPT-5.2	34	0.544	2.971	2.951	3.000	0.873	37.824

Beyond optimizing a single clinical policy, representing clinical decision-making as an observable trajectory creates an opportunity for studying variation among expert reasoning policies. Two policies can arrive at the same treatment recommendation while implementing meaningfully different information-acquisition and reasoning trajectories. Conversely, a policy may acquire appropriate evidence and reason coherently yet ultimately select a different but clinically defensible treatment. The absence of a unique treatment trajectory also has implications for how clinical reasoning should be modeled.

A retrospective dataset aggregated across clinicians do not necessarily represent samples from a single expert policy. Instead, it may represent a mixture of distinct clinician practice phenotypes (Krell et al., 2015). We use clinician practice phenotype to denote a recurrent pattern in how a clinician or clinician subgroup acquires information, interprets clinical evidence, and selects treatment actions across comparable clinical states. For example, clinicians may differ systematically in how much evidence they acquire before committing to treatment, or in how they trade treatment intensity against toxicity for clinically similar patients. Therefore, increasing the size of a clinical dataset can increase not only the number of patient states observed but also the number of expert reasoning policies represented in the data. A clinical policy trained such trajectories may consequently encounter heterogeneous reasoning patterns from clinicians with different but clinically defensible policies. As a preliminary analysis of physician-associated variation detailed in Appendix H, we find that the RL-trained policy does not show detectable heterogeneity across the chosen physician-associated subgroups, still, the results do not establish clinician practice phenotypes.

However, physician identity alone is insufficient to establish a clinician practice phenotype. Because each oncologist treated a different set of patients, differences between physician subgroups remain confounded by patient case mix and cannot be attributed specifically to differences in clinical decision policies. Establishing clinician practice phenotypes would require comparing information acquisition, reasoning, and treatment actions across comparable clinical states while accounting for differences in the patients encountered by each clinician. The trajectory structure captured by the environment provides the observables required for such an analysis.

Limitations and Future Work. Due to resource constraint, we select the deterministic mediator based on an embedding model as opposed to generative language models. The deterministic mediator cannot fully account for the variability in realistic clinical communications, posing a limitation on how faithful the environment capture the nature of the clinical interaction. The patient representations reconstructed from the clinical records are an approximation of the actual patient encounters specifically with gastrointestinal medical oncologists. In real world, the complexity of comprehensive patient records can span across temporality, conditions, specialties, hospital systems, and so forth. Additionally, we only conduct a single RL training run per training configuration due to resource constraint. We will continue working on identifying and modeling specific phenotypes of clinical practice, extending the clinical tasks, including treatment modifications, beyond first-line treatment selection, and improving the fidelity of the patient representation setup to the realistic clinical environment. We will also keep improving the performance of the mediator by, for example, assigning the mediator an adversarial objective and co-training it and turning the protocol into a two-player stochastic game.

AI USE STATEMENT

In this work, we used generative AI tools for aiding writing, discovery for related work, and providing feedback. We have not used generative AI tools for ideation and developing conceptual framework and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for editing software program. We have reviewed all AI-assisted work. We check LLM-generated content and software programs through manual verifications. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The aforementioned clinical database contains wholly de-identified medical records; no contact with human subjects occurred, and the study was therefore not considered human subjects research and was exempt from Institutional Review Board approval.

REPRODUCIBILITY STATEMENT

To comply with the data privacy, we are not able to provide the dataset readily available. However, the anonymized codebase for the data pre-processing, model training and evaluation, is available here.

REFERENCES

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. *arXiv [cs.CL]*, May 2025.

David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Nassir Navab, and Matthias Keicher. Language agents for hypothesis-driven clinical decision making with reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7vHUQCMAzG.

Li-Ching Chen, Travis Zack, Arda Demirci, Madhumita Sushil, Brenda Miao, Corynn Kasap, Atul Butte, Eric A. Collisson, and Julian C. Hong. Assessing large language models for oncology data inference from radiology reports. *JCO Clinical Cancer Informatics*, (8):e2400126, 2024. doi: 10.1200/CCI.24.00126. URL https://ascopubs.org/doi/abs/10.1200/CCI.24.00126. PMID:.

Li-Ching Chen, Travis Zack, Divneet Mandair, Aditya Mahadevan, Arvind Suresh, Yuta Ishiyama, Yiping Li, Julian C. Hong, and Atul J. Butte. From zero-shot to bedside: A practical playbook for adapting open-source large language models to clinical symptom extraction. In Peniel Argaw, Haoran Zhang, Sarah Jabbour, Payal Chandak, Jerry Ji, Sumit Mukherjee, Olawale Salaudeen, Trenton Chang, Elizabeth Healey, Fabian Gröger, Amin Adibi, Stefan Hegselmann, Benjamin Wild, and Ayush Noori (eds.), *Proceedings of the Fifth Machine Learning for Health Symposium*, volume 297 of *Proceedings of Machine Learning Research*, pp. 1023–1046. PMLR, 13–14 Dec 2026. URL https://proceedings.mlr.press/v297/chen26a.html.

Christopher Chiu, Silviu Pitis, and Mihaela van der Schaar. Simulating viva voce examinations to evaluate clinical reasoning in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=FEVfIPMy5b.

Yichun Feng, Jiawei Wang, Lu Zhou, Zhen Lei, and Yixue Li. DoctorAgent-RL: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16952–16956. IEEE, May 2026.

Shuohao Gao, Xuanzhong Chen, Lingxiao Luo, Zilin Ding, Rong Han, Rui Jiang, and Ting Chen. Salus: Strategic diagnostic testing for complex diagnosis via multi-agent reinforcement learning. In *Forty-third International Conference on Machine Learning*, June 2026.

Boyang Gu, Hongjian Zhou, Bradley Max Segal, Jinge Wu, Zeyu Cao, Hantao Zhong, Lei Clifton, Fenglin Liu, and David A. Clifton. Clinical-r1: Empowering large language models for faithful and comprehensive reasoning with clinical objective relative policy optimization. In Junde Wu, Jiazhen Pan, Jiayuan Zhu, Luyang Luo, Yitong Li, Min Xu, Yueming Jin, and Daniel Rueckert (eds.), *Proceedings of The Second AAAI Bridge Program on AI for Medicine and Healthcare*, volume 317 of *Proceedings of Machine Learning Research*, pp. 117–126. PMLR, 20–21 Jan 2026. URL https://proceedings.mlr.press/v317/gu26a.html.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=c1bTcrDmt4.

Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Do we need to verify step by step? rethinking process supervision from a theoretical perspective. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4BfaPHfhJ0.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. MedAgentBench: A virtual EHR environment to benchmark medical LLM agents. *NEJM AI*, 2(9), August 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, 11(14):6421, July 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.

Zahir Kanjee, Byron Crowe, and Adam Rodman. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*, 330(1):78–80, July 2023.

Robert W Krell, Scott E Regenbogen, and Sandra L Wong. Variation in hospital treatment patterns for metastatic colorectal cancer: Metastatic colorectal cancer treatment. *Cancer*, 121(11):1755–1761, June 2015.

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vQGHTyL0Jw.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. Mediq: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=W4pIBQ7bAI.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Y1rHxA58If.

Valentin Liévin, Samuel Schmidgall, Tim Strother, Alex Bijamov, Akshay Goel, Anil Palepu, Chunjong Park, Vahid Balazadeh, Min Woo Sun, Marius Guerard, Justin Chen, Dave Steiner, Vikram Dhillon, Ibrahim Azar, Akhil Mehta, Nicholas Spetsieris, Shilpan Shah, Maen Abdelrahim, Amit Dahiya, Yun Liu, Katherine Chou, Yossi Matias, Avinatan Hassidim, Dale R Webster, Quoc V Le, Raia Hadsell, Joelle Barral, Carey Radebaugh, Aleksandra Faust, Shekoofeh Azizi, Mike Schaekermann, Po-Hsuan Cameron Chen, Tao Tu, David Racz, and Lin Yang. ResidencyRL: Reinforcement learning in simulated clinical environments. *arXiv [cs.AI]*, August 2026.

Mengxian Lyu, Cheng Peng, Tim Jang, Ang Li, Mengyuan Zhang, Ziyi Chen, Leighton Elliott, Tianshi Liu, Lidice Galindo, Chiranjeevi Sainatham, Oscar F Borja-Montes, Kaleb E Smith, Ying Zhang, Lichao Sun, Jiang Bian, Gloria Lipori, Duane A Mitchell, Elizabeth A Shenkman, Yi Guo, Thomas J George, and Yonghui Wu. An agentic generative large language model for treatment planning of colorectal cancer. *arXiv [cs.CL]*, August 2026.

Mingyu Derek Ma, Xiaoxuan Wang, Yijia Xiao, Anthony Cuturrufo, Vijay S Nori, Eran Halperin, and Wei Wang. Memorize and rank: Elevating large language models for clinical diagnosis prediction. In *GenAI for Health: Potential, Trust and Policy Compliance*, November 2024.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models. *arXiv [cs.CL]*, July 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

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 [cs.CL]*, October 2025a.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. Quantifying the reasoning abilities of LLMs on clinical cases. *Nat. Commun.*, 16(1):9799, November 2025b.

Arya S Rao, Kaiz P Esmail, Richard S Lee, Sharon Jiang, Bianca Arraiza Carlo, Jasleen Gill, Praneet Khanna, Ezra Kalmowitz, Basile Montagnese, Kimia Heydari, Qiao Jiao, Ethan Bott, Dan Nguyen, Grace Wang, Michael Hood, Adam B Landman, and Marc D Succi. Large language model performance and clinical reasoning tasks. *JAMA Netw. Open*, 9(4):e264003, April 2026.

Silas Ruhrberg Estévez, Dyke Ferber, Mihaela van der Schaar, and Jakob Nikolas Kather. How to benchmark medical AI agents. *PLoS Med.*, 23(7):e1005170, July 2026.

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 Digit. Med.*, 9(1):499, April 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y K Li, Y Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv [cs.CL]*, February 2024.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067): 442–450, June 2025.

Kiril Vasilev, Alexandre Misrahi, Eeshaan Jain, Phil F Cheng, Petros Liakopoulos, Olivier Michielin, Michael Moor, and Charlotte Bunne. MTBBench: A multimodal sequential clinical decision-making benchmark in oncology. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, October 2025.

A Data & Cohort Construction

We conducted this study using retrospective, longitudinal, de-identified electronic health records (EHR)s from a large academic medical center in the United States, integrating data from their Clinical Data Warehouse and the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Patients were first qualified using cancer-specific SNOMED CT and ICD-10-CM diagnosis codes for colon cancer in the OMOP datasets. The index date was defined as the earliest qualifying cancer diagnosis date. Adult patients with a qualifying cancer diagnosis were then linked to chemotherapy episodes recorded in the OMOP database using the relevant chemotherapy episode concept. This is to ensure we select patients that were receiving the entirety of their oncology care within the specific hospital system, as opposed to seeing the specific hospital system for consultation but receiving care elsewhere, which would make definitive treatment planning more difficult to ascertain.

We consider an episode eligible for subsequent note collection if it was linked to a patient in the cancer cohort, had the predefined concept code for the cancer registry, and had an episode source concept name indicating chemotherapy. Among patients with eligible chemotherapy episodes, we collected unstructured clinical notes documented from 60 days before the index date through the chemotherapy episode date and retained unstructured notes from GI Medical Oncology encounters associated with clinicians whose title was MD.

The resulting notes were first processed into chronologically concatenated notes. These concatenated notes were then processed by GPT-5.2 to extract the treatment plan administered by the oncologist, the date on which the plan was documented, the clinical-trial enrollment, the rationale for the treatment decision, and the clinical factors informing the plan, including patient status, imaging, laboratory findings, and biomarkers. This processing extracts two classes of treatment plans, those which are eligible and those which are ineligible. The prompt for extracting the treatment plan and the reasoning behind the treatment plan is specified in Appendix J.1.

We consider treatment plans eligible if first, a definitive treatment was documented. By definitive, we mean that even though there may be many clinically defensible treatment plans, the clinician selected a single treatment plan based on their own latent reasoning. Second, the extracted treatment plan should have explicit documentation of the medication or treatment regimen. Lastly, the patient was not identified as being enrolled in a clinical trial, where decision would be based on completely unidentifiable variables, such as trial and slot availability at time of encounter. After an eligible treatment plan was identified and filtered for, we reconstructed the patient's clinical trajectory at the time of treatment planning. We implemented this reconstruction by first collecting GI Medical Oncology, MD-authored notes documented from 60 days before, including the treatment-plan date and then we used GPT 5.2 to process the per-patient concatenation of the notes to generate a structured record which we call the patient representation.

Clinical observations were therefore irregularly sampled according to the timing and frequency of EHR documentation, without temporal resampling or interpolation. Demographic characteristics and longitudinal diagnosis history were incorporated into the resulting patient representation. Missing values were retained without imputation to reflect nature of the EHR. No explicit administrative or clinical censoring was applied; the observation windows were defined relative to the cancer index date, chemotherapy episode date, and treatment-plan date.

The patient representation is a structured dictionary comprising three components (see Appendix J.2). First, a one-line summary of the history of present illness, representing the information known at the opening of the oncology consultation and serving as the initial observation. Second, an oncologic profile, including Eastern Cooperative Oncology Group (ECOG) performance status, disease staging, histology, biomarkers, and prior therapies, and, lastly, physiological data, including body measurements, vital signs, systemic examination findings, and medication history (Chiu et al., 2026).

The final dataset is partitioned at the eligible patient level into disjoint training, validation, and test sets, comprising 665, 143, and 142 patients, respectively, for a total of 950 patients.

B PATIENT REPRESENTATION AND ENVIRONMENT STATE

We will denote the i-th patient by a superscript i. Multiple trajectories g may be sampled for the same patient i. Hence, trajectory-dependent objects will additionally carry a superscript g. We will use $k \in \{0, \ldots, K-1\}$ for the RL environment step, t for token position within an agent-generated response, and n for the optimizer step. In our setup, we will call the RL environment step, a turn. We will call K, the budget.

We distinguish the underlying clinical state of the patient χ^i, the reconstructed patient representation ρ^i, the internal environment state $s_k^{i,g}$, and the observation $o_k^{i,g}$. The underlying clinical state $\chi^i \in \mathcal{X}_{\text{clin}}$ denotes the complete clinical condition of patient i, including clinically relevant properties whether or not they were measured or documented in the EHR. No clinical time elapses during a trajectory. This is an assumption about the simulated RL environment rather than a claim that the patient is static in the real clinical time. The agent's actions do not change the patient's physiology or generate counterfactual treatment outcomes. As a result, the nontrivial information dynamics arise from the sequential information acquisition, clinical inference and decision planning that the clinical agent takes.

The implemented environment therefore operates on a reconstructed patient representation $\rho^i = \rho\left(\ell^i, \mathcal{R}^i\right)$ where ℓ^i is a one-line patient summary available at the start of the episode and $\mathcal{R}^i$ is the patient-specific EHR record which is initially hidden but available through the retrieval. The construction of the patient representation that we explicitly described in the previous section refers to the mapping ρ.

To facilitate rapid RL training and minimize cost, we converted the unstructured clinical not documenting patient state and clinician decision to a structured key-value dictionary. The retrievable patient record is stored as a nested key-value dictionary. Let $\mathcal{F}^i$ denote clinical fields available for patient i. We write $\mathcal{R}^i = \left\{\left(\xi, v^i(\xi)\right) : \xi \in \mathcal{F}^i\right\}$, where ξ is a clinical field and $v^i(\xi)$ is its documented value. We use a deterministic flattening operation Flat recursively traverses the nested dictionary and separates each item within a list into an individual clinical field while keeping the contents of each list item together. The value $v^i(\xi)$ may be numerical or categorical and denotes the actual value that is kept by the patient reconstruction pipeline, not necessarily the information in the underlying clinical state χ^i and hence the patient representation ρ^i is inherently an incomplete and noisy proxy for the underlying clinical state. A clinical field is therefore a semantic address in the patient schema, e.g., age, performance status, histology or a specific record within the prior medical history. The value $v^i(\xi)$ may be numerical, categorical or structured and denotes the actual value that is kept by the patient reconstruction pipeline, not necessarily the information in the underlying clinical state χ^i and hence the patient representation ρ^i is inherently an incomplete and noisy proxy for the underlying clinical state.

The formulation of the partially observed Markov decision process will require us to know the fully observed RL trajectory $\omega^{i,g}$ which can be constructed from the fully observed state-action history $H_k^{i,g} = (s_0, a_0, s_0, a_1, ..., s_k)$, the patient representation ρ^i, the budget K and any additional information flow x that passes through the environment (parser and mediator) which the agent does not observe. The fully observed RL trajectory is then parametrized by

$$\omega_k^{i,g} = \left(H_k^{i,g}, \rho^i, K-k, x\right) \in \mathcal{S} \times \mathcal{A}, \tag{4}$$

where x denotes the rest of the environment information that goes to the judge model but the agent does not observe. We further remark that we will formulate the reward model as a terminal reward that does not appear within the history $H_k^{i,g}$ until the terminal step.

The agent, however, keeps track of the partially observed RL trajectory $\iota^{i,g}$ that it receives through observations $o_k^{i,g}$. It is constructed from the partial observation-action history $h_k^{i,g} = (o_0, a_0, o_1, a_1, ..., o_k)$ through an observation map Ω which filters out the patient representation and the environment information x; it parametrized by

$$\iota_k^{i,g} = \Omega\left(\omega_k^{i,g}\right) = \left(h_k^{i,g}, K-k\right) \in \mathcal{O} \times \mathcal{A}, \tag{5}$$

Both RL trajectories are subject to the initial conditions

$$H_0^{i,g} = s_0^{i,g} = \left(\ell^i, \rho^i\right), \quad h_0^{i,g} = \ell^i. \tag{6}$$

C Further Details on Agent Action Space, Action Parser and Frozen Patient-Record Mediator

C.1 Action Space

At the turn k, the clinical agent with a parameterized policy π_{θ_n} conditioned on the observed transcript generates a natural-language response

$$\mathcal{A} \ni a_k^{i,g} \sim \pi_{\theta_n}\left(\cdot \mid \iota_k^{i,g}\right). \tag{7}$$

Here, $a_k^{i,g}$ denotes the complete token sequence generated at one turn. Although the immediate action may contain free-form clinical reasoning at this stage, a structured action will be extracted by the parser.

C.2 Action Parser

To increase the learnability of the environment, we equip the clinical agent with a simple harness, which has a deterministic parser that maps the agent-generated response to an structured action

$$\bar{a}_k^{i,g} = \mathrm{Parse}\left(a_k^{i,g}\right) \in \mathcal{A}_{\text{investigate}} \cup \mathcal{A}_{\text{recommend}}. \tag{8}$$

The structured action space contains two action types. It may either investigate or recommend which terminates the episode. The action schema is organized by using three distinct separate tags: `<reasoning></reasoning>` for reasoning steps, `<query></query>` for clinical data acquisition or `<treatment_recommendation></treatment_recommendation>` for terminating the episode. If a response results in an investigation type action, it may contain multiple query tags. If no valid structured action can be extracted, then no clinical action is executed on the environment and the interaction turn is consumed. Thus, any malformed output will be penalized inherently by the environment. Each extracted query is additionally screened by deterministic regular expressions in order to suppress questions that directly ask the mediator to perform clinical inference as opposed to requesting the underlying documented variable e.g. a relevant biomarker. This restriction is imposed in order to softly define the separation between information acquisition phase and the clinical inference phase. When an investigation contains both accepted and rejected queries, accepted queries are executed independently, while corrective messages are returned for rejected queries. We will reserve $\bar{a}$ for the parsed action after its been processed by the regex.

C.3 Frozen Patient-Record Mediator

The mediator uses a Qwen3-Embedding-0.6B model whose parameters ϕ are not updated during the reinforcement learning and hence frozen. This model support query and document encoding, respectively. Before constructing the retrieval index, the nested patient record is flattened into a set of retrievable clinical fields. Each clinical field is then represented through two separate channels: a key channel containing the field path and the names of fields contained within the record, and a value channel containing the documented field content. The value-channel text is truncated to 300 characters for embedding only, while the full documented value is retained for the response.

Let us denote a natural-language string via ℓ. For each field $\xi \in \mathcal{F}^i$, let $d_{\mathrm{key}}(\xi)$ denote its key-channel representation and $d^i_{\mathrm{value}}(\xi)$ denote its value-channel representation. If we denote the query-embedding map as EmbQry_ϕ and the document-embedding map as EmbDoc_ϕ, then for any query ℓ, the mediator selects

$$\xi^\star\left(\ell;\rho^i\right) = \underset{\xi\in\mathcal{F}^i}{\arg\max}\max\left\{\frac{\left\langle \mathrm{EmbQry}_\phi(\ell), \mathrm{EmbDoc}_\phi\left(d_{\mathrm{key}}(\xi)\right)\right\rangle}{\left\|\mathrm{EmbQry}_\phi(\ell)\right\|_2 \left\|\mathrm{EmbDoc}_\phi\left(d_{\mathrm{key}}(\xi)\right)\right\|_2}, \frac{\left\langle \mathrm{EmbQry}_\phi(\ell), \mathrm{EmbDoc}_\phi\left(d^i_{\mathrm{value}}(\xi)\right)\right\rangle}{\left\|\mathrm{EmbQry}_\phi(\ell)\right\|_2 \left\|\mathrm{EmbDoc}_\phi\left(d^i_{\mathrm{value}}(\xi)\right)\right\|_2}\right\}. \tag{9}$$

where $||\cdot||$ is the Euclidean L_2 norm and the value-channel similarity is excluded for fields without a documented value. The mediator $\mathcal{M}_\phi$ then returns both the selected field and its full documented value

$$\mathcal{M}_\phi\left(\ell;\rho^i\right) = \left(\xi^\star\left(\ell;\rho^i\right), v^i\left(\xi^\star\left(\ell;\rho^i\right)\right)\right). \tag{10}$$

As a result, if we have an investigation action $\bar{a}$, the queries are ordered but passed as a batch through the mediator, the corresponding observation state at the next turn is then returned as

$$o_{k+1}^{i,g} = \text{Concatenate}\left(\left[\mathcal{M}_\phi\left(\ell;\rho^i\right) : \ell \in \left(\bar{a}_k^{i,g}\right)\right]\right), \tag{11}$$

and the partial history is updated as

$$\iota_{k+1}^{i,g} = \iota_k^{i,g} \oplus \left(a_k^{i,g}, o_{k+1}^{i,g}\right). \tag{12}$$

D Policy Optimization Algorithms

GRPO For GRPO, we optimize the clinical policy such that it maximizes the expected terminal return over the training distribution of patient representations which evaluates to

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{i\sim\mathcal{D}_{\text{train}}}\,\mathbb{E}_{\omega^{i,g}\sim\mathbb{P}_\theta}\left[r_{\text{clin}}^{i,g}\right]. \tag{13}$$

Here $\mathbb{P}_\theta$ is the trajectory measure induced by the environment induced jointly by the clinical policy and the environment transition rules.

The patient-specific group mean reward with penalty is

$$\bar{r}_{\text{GRPO}}^{i} = \frac{1}{G}\sum_{g=1}^{G} r_{\text{clin}}^{i,g}. \tag{14}$$

The group relative advantage for each rollout g is

$$\widehat{\text{Adv}}_{\text{GRPO}}^{i,g} = \frac{r_{\text{GRPO}}^{i,g} - \bar{r}_{\text{GRPO}}^{i}}{\text{std}_g\left[r_{\text{GRPO}}^{i,g}\right] + \epsilon_{\text{num}}}. \tag{15}$$

GDPO For GDPO, the relative advantage is defined as follows

$$\begin{aligned}
\widehat{\text{Adv}}_{\text{info}}^{i,g} &= \frac{r_{\text{info}}^{i,g} - \text{mean}_{g'=1,\ldots,G}\left[r_{\text{info}}^{i,g'}\right]}{\text{std}_{g'=1,\ldots,G}\left[r_{\text{info}}^{i,g'}\right] + \epsilon_{\text{num}}},\\
\widehat{\text{Adv}}_{\text{regimen}}^{i,g} &= \frac{r_{\text{regimen}}^{i,g} - \text{mean}_{g'=1,\ldots,G}\left[r_{\text{regimen}}^{i,g'}\right]}{\text{std}_{g'=1,\ldots,G}\left[r_{\text{regimen}}^{i,g'}\right] + \epsilon_{\text{num}}},\\
\widehat{\text{Adv}}_{\text{reason}}^{i,g} &= \frac{r_{\text{reason}}^{i,g} - \text{mean}_{g'=1,\ldots,G}\left[r_{\text{reason}}^{i,g'}\right]}{\text{std}_{g'=1,\ldots,G}\left[r_{\text{reason}}^{i,g'}\right] + \epsilon_{\text{num}}},\\
\widehat{\text{Adv}}_{\text{complete}}^{i,g} &= \frac{\mathbb{1}_c^{i,g} - \text{mean}_{g'=1,\ldots,G}\left[\mathbb{1}_c^{i,g'}\right]}{\text{std}_{g'=1,\ldots,G}\left[\mathbb{1}_c^{i,g'}\right] + \epsilon_{\text{num}}},\\
\widehat{\text{Adv}}_{\text{parse}}^{i,g} &= \frac{\mathbb{1}_p^{i,g} - \text{mean}_{g'=1,\ldots,G}\left[\mathbb{1}_p^{i,g'}\right]}{\text{std}_{g'=1,\ldots,G}\left[\mathbb{1}_p^{i,g'}\right] + \epsilon_{\text{num}}}.
\end{aligned} \tag{16}$$

Equal criterion weights are used in our experiments and thus we simply sum all coefficients up

$$\widetilde{\text{Adv}}_{\text{GDPO}}^{i,g} = \widehat{\text{Adv}}_{\text{info}}^{i,g} + \widehat{\text{Adv}}_{\text{regimen}}^{i,g} + \widehat{\text{Adv}}_{\text{reason}}^{i,g} + \widehat{\text{Adv}}_{\text{complete}}^{i,g} + \widehat{\text{Adv}}_{\text{parse}}^{i,g}.$$

GDPO also applies a second normalization over all rollouts in the optimizer batch.

$$\widehat{\text{Adv}}_{\text{GDPO}}^{i,g} = \frac{\widetilde{\text{Adv}}_{\text{GDPO}}^{i,g} - \text{mean}_{(i,g)\in B\times\{1,\ldots,N_g\}}\left[\widetilde{\text{Adv}}_{\text{GDPO}}{}^{i,g}\right]}{\text{std}_{(i,g)\in B\times\{1,\ldots,N_g\}}\left[\widetilde{\text{Adv}}_{\text{GDPO}}\right] + \epsilon_{\text{num}}}. \tag{17}$$

Objective Let $x_{k,t}^{i,g}$ be the agent-generated token at environment step k and token position t and let $\iota_{t,k}$ be the policy context immediately before that token. Then the likelihood ratio between the current policy and the behavior-policy is

$$w_{k,t}^{i,g}(\theta) = \frac{\pi_{\theta(n)|A}\left(x_{k,t}^{i,g} \mid \iota_{k,t}^{i,g}\right)}{\pi_{\bar{\theta}(n)|A}\left(x_{k,t}^{i,g} \mid \iota_{k,t}^{i,g}\right)}. \tag{18}$$

Let $\widehat{\mathrm{Adv}}^{i,g}$ denote either the GRPO coefficient or the GDPO coefficient. Because the reward is only available after termination, we broadcast this to every agent-generated token in the rollout, the corresponding clipped token-level contribution

$$\mathcal{L}_{k,t}^{i,g}(\theta) = \min\left[w_{k,t}^{i,g}(\theta)\widehat{\mathrm{Adv}}^{i,g}, \mathrm{clip}\left(w_{k,t}^{i,g}(\theta), 1-\varepsilon, 1+\varepsilon\right)\widehat{\mathrm{Adv}}^{i,g}\right]. \tag{19}$$

Only tokens generated by the clinical agent contribute to the policy-gradient term. Tokens belonging to the system prompt, initial patient observation, parser messages, or mapper responses are masked. The clipped objective is then

$$J_{\text{clip}}(\theta) = \mathbb{E}\left[\frac{1}{N_g}\sum_{g=1}^{N_g}\frac{1}{N_{\text{tok}}^{i,g}}\sum_{k,t}\mathbb{1}_{\textbf{agent}}(k,t)\mathcal{L}_{k,t}^{i,g}(\theta) - \beta_{\mathrm{KL}}D_{\mathrm{KL}}\left(\pi_{\theta(n)|A}\|\pi_{\text{ref}}^{|A}\right)\right], \tag{20}$$

where the reference policy π_{ref} is the frozen initial clinical-agent policy.

E Experimental Setup

We train Llama-3.1-8B-Instruct using parameter-efficient policy optimization under a five-turn interaction budget. At each turn, the agent can generate at most 512 new tokens. We generate a group of eight trajectories for each patient. We use AdamW optimizer with a constant learning rate of 1×10^{-6}. We enforce Kullback–Leibler (KL) divergence to regularize the update to the new policy against the reference policy with coefficient 0.01. We apply low rank adaptation (LoRA) to train the agent policy. We attach LoRA adapters to all attention and feed-forward projection matrices, i.e., query, key, value, output, gate, up, and down projections, of the agent policy, while the rest of the network is frozen and loaded with 4-bit quantization. We set the LoRA rank to 32, scaling factor α to 32, and dropout to 0.

The environment uses a frozen Qwen3-Embedding-0.6B mediator and GPT-5.2 judge throughout training and evaluation. We select GPT-5.2 because it is the most performant model available through our institutional secure endpoint, and set temperature to 0 with up to three retries. The context window is by default 400,000 tokens with maximum output 128,000 tokens per response. The judge returns a dictionary containing the numeric score for each reward axis together with natural-language grading rationale.

During the training, we use the complete trajectory to update the agent policy. The terminal reward is broadcast to all tokens generated by the agent policy for gradient update, while the tokens that are not generated by the agent, including system instructions and mediator responses, are masked from the gradient update. We assign equal weight to each reward axis during reward scalarization. To encourage diversity in the rollout, we use sampling-based decoding with temperature 0.6 and top-p 0.95. We process each of the 665 training patients once, so training consists of 665 patient-level steps. Within each patient-level step, corresponding trajectory group may undergo multiple inner optimization steps. We checkpoint the policy every 50 patient-level steps. For each training configuration, we independently select the checkpoint with the highest Mean@1 total scalar reward on the validation set with temperature set 0 and report test-set performance using the selected checkpoint. The test set is not used for checkpoint selection.

For both GRPO and GDPO, each patient contributes one group of eight sampled trajectories, which we reuse for three inner optimization steps before proceeding to the next patient. The inner optimization steps allows reusing the already sampled trajectories for multiple gradient updates. We optimize the PPO clipped surrogate with $\epsilon = 0.2$ to bound how far each update moves the policy.

F Additional Figures

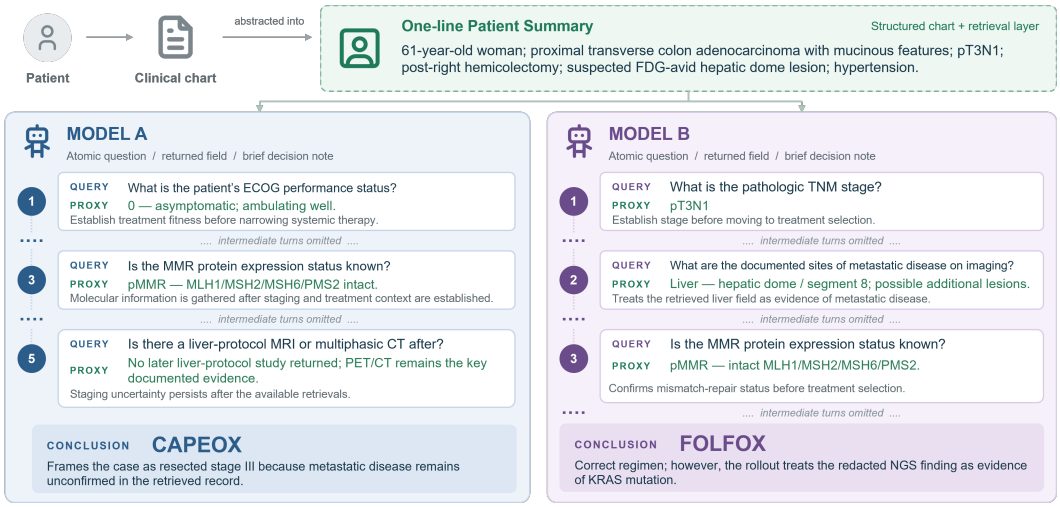
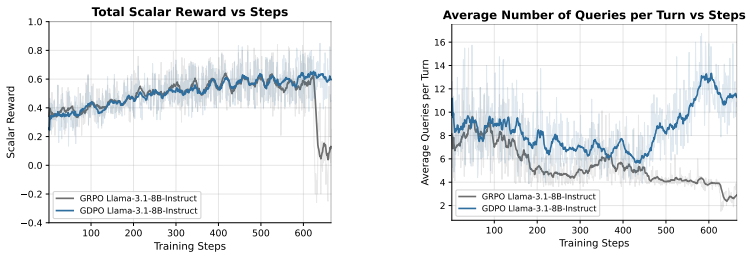


Figure 3: Illustrative example of trajectories from the clinical agent given a patient case. The clinical agent interacts the patient environment in order to retrieve relevant information necessary for selecting the treatment. In the process, they will describe the reasoning behind the retrieval certain information, and the inference on the retrieved information.



(a) Total scalar reward.

(b) Average number of queries.

Figure 4: Learning dynamics of GRPO and GDPO with Llama-3.1-8B-Instruct. We report the total scalar reward and the average number of queries issued by the policy. Dark curves show a 15-step moving average and faint curves show the corresponding unsmoothed training trajectory. Each curve corresponds to a single training run.

G Reinforcement Learning using Different Policy Backbones

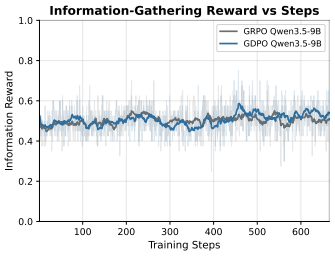
We investigate the efficacy of the reinforcement learning (RL) environment setup across different policy backbones, including Llama-3.1-8B-Instruct, Qwen-3.5 9B and GLM-4-9B-0414. RL improves performance on different policy backbones, although the magnitude of improvement depends strongly on the backbone. For Qwen3.5-9B, GRPO and GDPO increase Mean@3 from -0.0142 for the base policy to 0.1994 and 0.3429, respectively. For GLM-4-9B-0414, the Mean@3 improves from -0.0163 to 0.5313 and 0.5105 after GRPO and GDPO training, respectively. Both policy backbones demonstrate improvement in each reward axes during training, as shown in Figures 5 and 6.

H Quantitative Analysis of Information-Acquisition Behavior

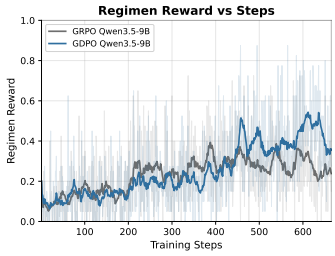
We quantify the behavior change in information acquisition before and after reinforcement learning and the difference between different policies. Using the trajectories collected from running the

Table 4: Performance on each reward axis. Information, treatment, and reasoning are scored from 0–5; completeness and format are scored from 0–1. Best results are shown in **bold** and second-best results are underlined.

		Information (0–5)	Regimen (0–5)	Reasoning (0–5)	Complete (0–1)	Format (0–1)
GRPO Llama-3.1-8B-Instruct	Mean@3	2.754	3.366	2.688	0.960	1.000
	Best@3	3.000	3.789	2.944	1.000	1.000
GDPO Llama-3.1-8B-Instruct	Mean@3	2.953	**3.453**	2.791	**0.988**	1.000
	Best@3	3.176	**3.845**	3.077	1.000	1.000
GRPO Qwen3.5-9B	Mean@3	2.606	1.453	2.298	0.439	1.000
	Best@3	3.007	2.796	2.768	0.824	1.000
GDPO Qwen3.5-9B	Mean@3	2.725	2.033	2.399	0.664	1.000
	Best@3	3.063	3.183	2.852	0.951	1.000
GRPO GLM-4-9B-0414	Mean@3	2.728	2.967	2.444	0.972	1.000
	Best@3	3.056	3.556	2.859	0.993	1.000
GDPO GLM-4-9B-0414	Mean@3	2.629	2.854	2.455	0.953	1.000
	Best@3	2.958	3.521	2.901	1.000	1.000
GPT-5.2	Mean@3	**2.977**	2.702	**3.016**	0.831	1.000
	Best@3	**3.232**	3.380	**3.352**	0.972	1.000
Llama-3.1-8B-Instruct	Mean@3	2.124	1.876	1.927	0.859	1.000
	Best@3	2.458	2.746	2.310	1.000	1.000
Qwen3.5-9B	Mean@3	2.406	0.437	2.085	0.143	1.000
	Best@3	2.852	1.063	2.338	0.338	1.000
GLM-4-9B-0414	Mean@3	2.002	0.678	1.723	0.225	1.000
	Best@3	2.394	1.542	2.070	0.500	1.000



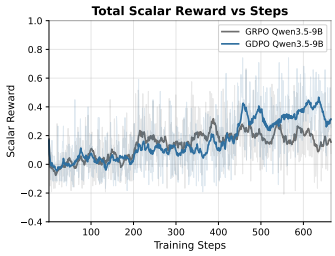
(a) Information-gathering reward.



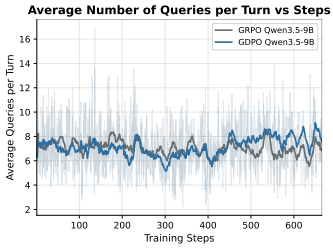
(b) Treatment reward.



(c) Reasoning reward.



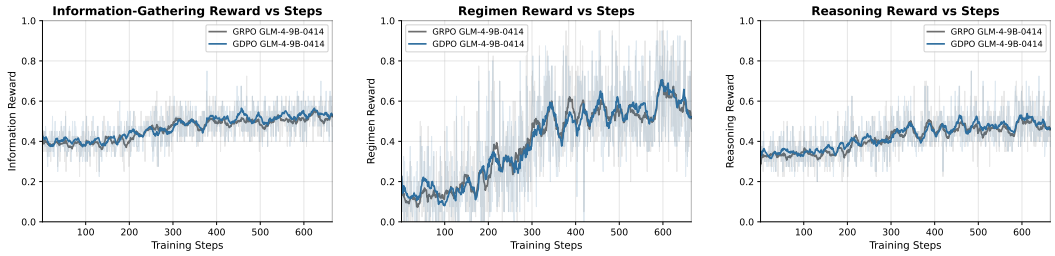
(d) Total scalar reward.



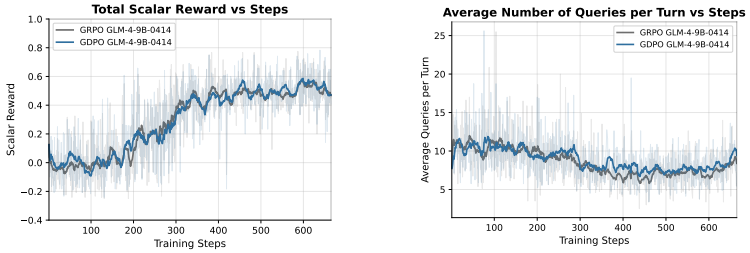
(e) Average number of queries.

Figure 5: GRPO and GDPO training dynamics on Qwen3.5-9B policy backbone. Dark curves show a 15-step moving average and faint curves show the corresponding unsmoothed training trajectory. Each curve corresponds to a single training run.

held-out test set ($n = 142$), we extract the number of queries the policy issues, the number of unique clinical field retrieved by the mediator, the number of terminal turn, the numeric score for information-gathering on a scale from 0 to 5, and calculate the rate of redundant queries. We first average over 3 sampled trajectories within each patient, and then conduct the statistical comparison



(a) Information-gathering reward. (b) Treatment reward. (c) Reasoning reward.



(d) Total scalar reward. (e) Average number of queries.

Figure 6: GRPO and GDPO training dynamics on GLM-4-9B-0414 policy backbone. Dark curves show a 15-step moving average and faint curves show the corresponding unsmoothed training trajectory. Each curve corresponds to a single training run.

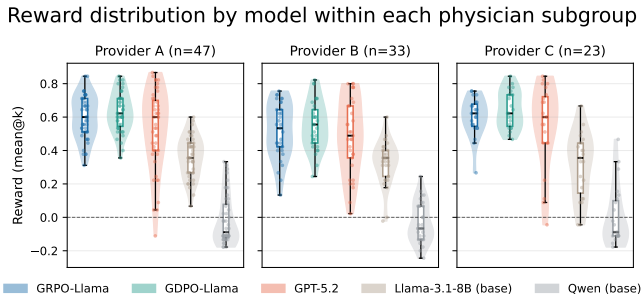
between policies in pair per patient. We report the mean paired different with a 95% bootstrap confidence interval, the median, and a paired Wilcoxon signed-rank p.

Compared to base Llama, GRPO reduces the number of queries by 4.65 per case and the number of unique fields retrieved by 2.09, while increasing information-gathering reward by 0.63. Conversely, GDPO increases the number of queries by 2.22 per case and the number of unique fields retrieved by 2.65, while increasing information-gathering reward by 0.83. Compared to GPT-5.2, both RL-trained Llama polices issue approximately 27 to 34 fewer queries per case and retrieve approximately 10 to 14 fewer unique fields than GPT-5.2.

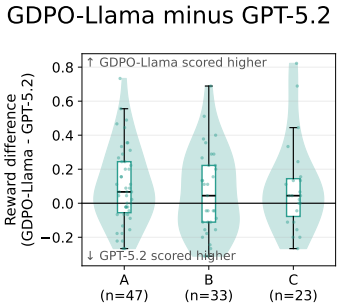
Together, these results provide quantitative evidence that the learned policy adapts its interaction behavior to the structure of the reinforcement learning environment. The information-gathering reward rubric favors acquisition of variables relevant to the decision while penalizing unnecessary or redundant investigation, including continued information gathering after sufficient evidence is available.

Table 5: Patient-level changes in information-gathering behavior on the held-out test set. Values report the mean paired difference between each RL-trained policy based on Llama-3.1-8B-Instruct and the reference policy, with 95% bootstrap confidence intervals.

Policy	Queries/case	Unique fields	Information Reward
Llama-3.1-8B (Base)	8.98 [8.42, 9.57]	6.21	2.12
GRPO-Llama	4.33 [4.17, 4.51]	4.12	2.75
GDPO-Llama	11.20 [10.84, 11.57]	8.86	2.95
Qwen (Base)	7.26 [6.87, 7.65]	5.20	2.41
GRPO-Qwen	6.96 [6.66, 7.25]	5.18	2.61
GDPO-Qwen	7.96 [7.57, 8.35]	5.73	2.73
GLM-4-9B (Base)	10.40 [9.81, 11.03]	5.23	2.00
GRPO-GLM	8.07 [7.69, 8.47]	6.12	2.73
GDPO-GLM	8.19 [7.76, 8.64]	6.04	2.63
GPT-5.2	38.41 [36.57, 40.27]	18.98	2.98



(a) Per-patient reward distributions across policies within the three largest primary-oncologist subgroups.



(b) Paired per-patient reward difference between the specialized GDPO policy and GPT-5.2.

Figure 7: Policy performance across primary-oncologist subgroups. **(a)** Per-patient reward distributions for each policy. **(b)** Per-patient reward difference between the specialized GDPO policy and GPT-5.2. Positive values indicate higher reward for GDPO.

Table 6: Paired comparisons of policy reward against GPT-5.2 within physician-associated patient subgroups. Differences are computed per patient using Mean@3 scalar reward. Confidence intervals are 95% bootstrap confidence intervals for the mean paired difference. p-values are from paired Wilcoxon signed-rank tests with Holm correction within each provider subgroup.

Provider	Policy	n	Mean Δ	95% CI	Median Δ	Holm p
A	GRPO Llama	47	0.071	[0.009, 0.137]	0.044	0.109
	GDPO Llama	47	0.100	[0.036, 0.169]	0.067	**0.018**
	Llama base	47	-0.165	[-0.239, -0.087]	-0.200	7.6×10^{-4}
	Qwen base	47	-0.544	[-0.619, -0.463]	-0.556	1.6×10^{-8}
B	GRPO Llama	33	0.032	[-0.055, 0.124]	0.000	0.674
	GDPO Llama	33	0.063	[-0.015, 0.149]	0.044	0.396
	Llama base	33	-0.158	[-0.231, -0.082]	-0.133	**0.002**
	Qwen base	33	-0.519	[-0.593, -0.444]	-0.556	2.2×10^{-6}
C	GRPO Llama	23	0.055	[-0.045, 0.165]	0.044	0.770
	GDPO Llama	23	0.096	[-0.002, 0.210]	0.044	0.422
	Llama base	23	-0.238	[-0.375, -0.092]	-0.178	**0.014**
	Qwen base	23	-0.547	[-0.664, -0.421]	-0.622	1.6×10^{-4}

Δ denotes policy reward minus GPT-5.2 reward. Bold p-values indicate $p < 0.05$ after Holm correction.

I Exploratory Analysis Across Primary Oncologist Subgroups

To examine whether policy performance is consistent across physician-associated patient subgroups, we stratified the held-out test set according to the patient's primary oncologist. The primary oncologist was identified from the author of the clinical note from which the documented treatment plan was extracted. We selected the three oncologists with the largest number of patients in the test set, corresponding to 47, 33, and 23 patients, respectively. We then compared the per-patient reward distributions of the evaluated policies within each subgroup.

The GDPO-trained Llama policy has a positive median reward difference relative to GPT-5.2 in all three physician subgroups, with median paired differences of 0.067, 0.044, and 0.044 for Providers A, B, and C, respectively (see Figure 7 and Table 6). The difference reached statistical significance in Provider A (mean paired difference 0.100, 95% CI $[0.036, 0.169]$, Holm-corrected $p = 0.018$) but not in Providers B or C (Holm-corrected $p = 0.396$ and $p = 0.422$, respectively). We found no evidence that the GDPO–GPT-5.2 reward difference varied across physician subgroups (Kruskal–Wallis $H = 0.464$, $p = 0.793$). Absolute reward distributions varied across physician subgroups for the GDPO policy (Kruskal–Wallis $H = 8.585$, $p = 0.014$; see Table 7), although these comparisons are confounded by differences in patient case mix.

Table 7: Per-patient Mean@3 scalar reward across physician-associated patient subgroups. Kruskal–Wallis comparisons are descriptive because provider subgroups may differ in patient case mix.

Policy	Provider A	Provider B	Provider C	H	p
GRPO Llama	0.599	0.519	0.604	5.400	0.067
GDPO Llama	0.627	0.550	0.644	8.585	**0.014**
GPT-5.2	0.528	0.487	0.549	1.784	0.410
Llama base	0.363	0.329	0.311	1.426	0.490
Qwen base	-0.017	-0.032	0.002	0.248	0.883

J Prompt Details

J.1 Instruction for extracting clinical decision making process

Clinician Decision Extraction Prompt

```
system_prompt = """
You are an Expert Oncology Data Abstractor and Clinical Logic
    Analyzer.

You will receive a chronological list of clinical notes for a single
     colon cancer patient.

Your task is to reconstruct the clinician's decision process in a
    structured and clinically grounded way.

PRIMARY GOALS:
1. Identify the first-line systemic therapy plan and the dose
    intensity used.
2. Identify whether the plan involved a clinical trial.
3. Reconstruct the decision pathway the clinician used to reach the
    plan, with emphasis on the sequence of information gathering,
    interpretation, and decision-making.

ABSTRACTION PRINCIPLES:
- Be specific and clinically concrete.
- Preserve chronology.
- Separate observations from actions and from conclusions.
- Do not collapse multiple facts into vague summaries.
- Do not infer facts that are not stated or clearly implied.
- If a detail is not documented, output "Not Explicitly Documented"
    or null as appropriate.
- Prefer the exact wording from the note when possible, but
    normalize medically equivalent phrasing into a standard label.

DECISION PATHWAY CATEGORIES:
You must extract the following categories separately:

1. PATIENT_FACTORS
   - Age
   - Sex
   - ECOG / performance status
   - Comorbidities
   - Symptoms relevant to treatment planning
   - Prior treatment history

2. PATHOLOGY
   - Biopsy result
   - Histology
   - Grade / differentiation
   - Tumor subtype
   - Margins
   - Lymphovascular invasion
   - Perineural invasion
```

```
   - Positive nodes on surgical pathology
   - Pathologic T/N stage if reported
   - Other pathology-derived risk features

3. IMAGING
   - Imaging studies ordered or reviewed
   - Modality and anatomic target
   - Key findings from each study
   - Whether the study was obtained for staging, restaging, or
    treatment planning
   - Imaging findings that changed management

4. STAGING
   - Clinical stage
   - TNM stage
   - Stage group
   - Localized vs locally advanced vs metastatic status
   - Resectability / operability
   - Sites of metastatic disease
   - Basis for the staging conclusion

5. BIOMARKERS
   - Molecular / genomic results
   - MSI / MMR status
   - KRAS / NRAS
   - BRAF
   - HER2
   - Other predictive or prognostic biomarkers
   - Any biomarker used to guide therapy selection

6. LABS_AND_SAFETY
   - Organ function
   - Hematologic values
   - Contraindications or safety constraints
   - Toxicity concerns relevant to regimen choice or dose adjustment

7. ACTIONS_TAKEN
   - Tests ordered
   - Consults requested
   - Tumor board discussion
   - Surgery referral
   - Treatment sequencing decisions
   - Follow-up plans

8. OPTIONS_WAIGHED
   - Alternate regimens discussed
   - Why each option was chosen, rejected, or deferred
   - Any uncertainty or pending data that influenced the decision

NARRATIVE REQUIREMENTS:
- The narrative_summary must be chronological.
- It must explicitly describe the sequence:
  1. What information was available initially
  2. What was ordered or reviewed next
  3. What conclusion those data supported
  4. How the treatment decision followed
- Use concrete clinical facts, not vague language like "based on the
    clinical picture."
- End the narrative with the final regimen and the dose intensity
    decision.
- If dose was reduced, explicitly state the documented reason.
- If dose intensity was not documented, state that clearly.

OUTPUT RULES:
- Return only valid JSON.
- Use null where a value is absent and "Not Explicitly Documented"
    only where a string is required.
- Do not add commentary outside the JSON.
```

```
"""

user_prompt = f"""
Here is the patient chronological history:
'''
{history_text}
'''
Please use the plan date of {plan_date} to help anchor the timeline
    if that information is available in the notes.
Extract the clinician's decision process into the following JSON
    structure exactly:

{{
  "treatment_plan": {{
    "regimen": "String",
    "is_clinical_trial": Boolean,
    "clinical_trial_details": "String or null",
    "regimen_reasoning": "String",
    "dose_intensity": "String",
    "dose_reasoning": "String or null"
  }},
  "plan_date": "String (YYYY-MM-DD)",
  "is_definitive_decision": Boolean,
  "decision_pathway": {{
    "patient_factors": [
      {{
        "variable": "String",
        "value": "String",
        "clinical_relevance": "String"
      }}
    ],
    "pathology": [
      {{
        "variable": "String",
        "value": "String",
        "clinical_relevance": "String"
      }}
    ],
    "imaging": [
      {{
        "study": "String",
        "finding": "String",
        "clinical_relevance": "String",
        "purpose": "String"
      }}
    ],
    "staging": [
      {{
        "variable": "String",
        "value": "String",
        "basis": "String"
      }}
    ],
    "biomarkers": [
      {{
        "variable": "String",
        "value": "String",
        "clinical_relevance": "String"
      }}
    ],
    "labs_and_safety": [
      {{
        "variable": "String",
        "value": "String",
        "clinical_relevance": "String"
      }}
    ],
    "actions_taken": [
```

```
      {{
        "action_type": "String",
        "action": "String",
        "purpose": "String"
      }}
    ],
    "options_weighed": [
      {{
        "regimen": "String",
        "status": "Chosen/Rejected/Discussed/Deferred",
        "reasoning": "String"
      }}
    ],
    "narrative_summary": "String"
  }},
  "extraction_metadata": {{
    "confidence_score": "Number 1-10",
    "reasoning_for_extraction": "String"
  }}
}}
"""
```

J.2 Instruction for building the patient representation

Patient Representation Prompt

```
json_template = """
{
    "clinical_summary_one_liner": "String (The Priming One-Liner)",

    "history_freetext": "Concise summary of the HPI",

    "oncology_profile": {
        "performance_status": {
        "ecog_score": Number (0-5) or null,
        "description": "String",
        "source_text": "Exact quote",
        "is_inferred": Boolean
        },
        "staging": {
        "tnm_clinical": "String (e.g., cT3N1M0)",
        "tnm_pathological": "String (e.g., pT3N2aM1a)",
        "stage_group": "String (e.g., 'Stage IIIB')",
        "staging_proxies": ["String"],
        "metastatic_sites": ["String"]
        },
        "histology": {
        "type": "String",
        "grade": "String",
        "primary_site": "String",
        "notes": "String"
        },
        "biomarkers": [
        {
            "marker": "String",
            "status": "String",
            "variant": "String",
            "date": "String"
        }
        ],
        "prior_therapies": [
        {
```

```
            "regimen": "String",
            "response": "String",
            "reason_stopped": "String",
            "date_last_dose": "String"
        }
        ]
    },

    "physical": {
        "body_measurements": {
        "height_cm": Number or null,
        "weight_kg": Number or null,
        "bsa_m2": Number or null,
        "weight_change": "String"
        },
        "vitals": {
        "blood_pressure_systolic": [Number],
        "blood_pressure_diastolic": [Number],
        "heart_rate": [Number],
        "temperature": Number or null,
        "oxygen_saturation": Number or null,
        "temporal_notes": {
            "heart_rate": ["String"],
            "blood_pressure_systolic": ["String"]
        }
        },
        "systems": {
        "neurological": {
            "neuropathy": {
                "present": Boolean,
                "grade": "String",
                "description": "String"
            },
            "mental_status": { "name": "Mental Status", "description
    ": String }
        },
        "abdominal": { "distension": Boolean, "tenderness": Boolean,
     "mass": Boolean },
        "general": { "cachexia": Boolean, "appearance": String }
        }
    },

    "investigations": {
        "blood": {
        "neutrophils_abs": { "name": "ANC", "value": Number, "units
    ": "cells/uL", "flag": String },
        "hemoglobin": { "name": "Hemoglobin", "value": Number, "flag
    ": String },
        "platelets": { "name": "Platelets", "value": Number, "flag":
     String },
        "bilirubin_total": { "name": "Total Bilirubin", "value":
    Number, "flag": String },
        "ast": { "name": "AST", "value": Number, "flag": String },
        "alt": { "name": "ALT", "value": Number, "flag": String },
        "creatinine": { "name": "Creatinine", "value": Number, "flag
    ": String },
        "cea": { "name": "CEA", "value": Number, "trend": "String"
    },
        "ca19_9": { "name": "CA 19-9", "value": Number }
        },
        "other": {
        "ecg": { "name": "ECG", "value": String }
        }
    },

    "imaging": {
        "reports": [
```

```
        { "modality": "String", "region": "String", "date": "String
    ", "report_text": "String", "key_findings": ["String"] }
        ]
    },

    "medications": [
        {
        "name": String,
        "dosage": String,
        "frequency": String,
        "route": String,
        "indication": String,
        "current": Boolean
        }
    ],

    "past_medical_history": [
        {
        "condition": "String",
        "present": "Boolean",
        "ongoing": "Boolean",
        "impact_on_treatment": "String"
        }
    ],
    "history": {
        "chief_complaint": "String",
        "symptoms": [
        {
            "name": "String",
            "present": "Boolean",
            "system": "String",
            "onset": "String",
            "primary": "Boolean"
        }
        ]
    },
    "allergies": [
        {
        "allergen": "String",
        "reaction": "String",
        "severity": "String"
        }
    ],
    "social_history": {
        "smoking_status": "String (e.g., Active, Former, Never)",
        "alcohol_use": "String",
        "living_situation": "String (e.g., Lives with spouse, Lives
    alone)"
    },
    "organ_function_scores": {
        "creatinine_clearance_or_egfr": "String or null",
        "child_pugh_score": "String or null"
    }
    "demographics": { "age": Number, "gender": String, "ethnicity":
    String },

    "diagnosis": [
        {
        "name": String,
        "icd_10": String,
        "date_diagnosis": "String (YYYY-MM-DD)",
        "source": "String (Injected or Notes)",
        "relevant_keys": ["JSON Path"]
        }
    ]
}
"""
system_instruction = f"""
```

```
You are an Expert Oncology Data Abstractor. Your goal is to
    structure clinical text for a downstream AI Treatment Planner.

### INPUT CONTEXT
You will be provided with:
1.  **Clinical Notes:** Unstructured text (HPI, Physical, Assessment
    ).
2.  **Structured Demographics (INJECTED):** Trusted source of truth
    for Age/Gender/Race.
3.  **Diagnosis History (INJECTED):** A list of past diagnoses with
    dates.

### CRITICAL LEAKAGE GUARDRAILS
1.  **NO FUTURE PLANS**: You must capture the patient's history and
    current status ONLY.
    * **IGNORE** sections labeled "PLAN", "RECOMMENDATION", or "
    ASSESSMENT & PLAN" regarding *new* orders.
    * **DO NOT** extract the treatment the doctor is *about* to
    prescribe or considering.
    * *Exception*: You MAY extract summaries of *past* treatments
    found in these sections.

### SPECIAL INSTRUCTIONS FOR ONCOLOGY
1.  **Data Hierarchy (Structured vs. Unstructured)**:
    * **Demographics**: Use the INJECTED structured data as the
    primary source. Only override if the text explicitly corrects a
    data error.
    * **Diagnosis**: Use INJECTED history to populate `
    past_medical_history`. However, use the Clinical Notes to
    determine the *active* oncology problem (e.g., if history says "
    Lung Nodule (2020)" but text says "Biopsy confirms
    Adenocarcinoma", prioritze the text for the `oncology_profile`).

2.  **ECOG Performance Status (Noise Handling)**:
    * **Rule**: Prioritize the score/description from the **latest
    visit**.
    * **Conflict Resolution**: If the note contains contradictory
    scores (e.g., "ECOG 0" in header, but "Patient is bedbound" in
    text), prioritize the **functional description** over the number
    .
    * **Inference**: If not explicitly stated, infer from latest
    physical ability (e.g., "active" = 0, "up >50% of day" = 1) and
    mark `is_inferred: true`.

3.  **Biomarkers (Deep Extraction)**:
    * **Granularity**: Extract specific variants (e.g., "V600E", "
    G12C") if available.
    * **Standardization**:
        * "No mutation" / "None" -> "Wild-type"
        * "MMR Proficient" / "Intact" -> "pMMR"
        * "Loss of MLH1/PMS2" -> "dMMR"
    * **Complex Cases**: If a patient is KRAS WT but BRAF V600E,
    list BOTH as separate entries.

4.  **Staging (Robustness)**:
    * Look for TNM scores (e.g., T3N1M0).
    * **Proxies**: If TNM is missing, extract literal text that
    determines intent (e.g., "Unresectable", "Widely metastatic", "
    Liver-limited").

5.  **Prior Cancer Therapy**:
    * Explicitly extract previous regimens in `oncology_profile.
    prior_therapies`.
    * **Crucial**: Capture `reason_for_discontinuation` (Progression
     vs Toxicity) to prevent re-prescribing failed drugs.
```

```
6. Biomarker Status Guardrails: If a biomarker is explicitly
    mentioned as ordered but not resulted, record its status as "
    Pending". If it is not mentioned at all, do not invent it.

7. Allergies & Contraindications: Aggressively scan for allergies,
    especially to common chemotherapy agents (e.g., Platinums,
    Fluorouracils) or supportive meds (e.g., Sulfa).
### THE "PRIMING" ONE-LINER
You must generate a field called 'clinical_summary_one_liner'.
* **Purpose**: To allow a downstream model to instantly recognize
    the cancer context AND safety constraints (comorbidities).
* **Guardrail**: Do not include the treatment plan in the summary.
    The downstream model is tasked with figuring out the treatment
    plan.
* **Format**: "[Age]yo [Gender] with [Stage] [Histology] of the [
    Primary Site], presenting with [Key Symptom/Status]. Notable for
     [Major Comorbidities]."
* **Comorbidity Logic**: Include conditions that impact chemotherapy
     safety (e.g., Diabetes, CKD, CHF, COPD, Autoimmune disease). Do
     not list minor issues (e.g., GERD, mild anxiety).
* **Example 1**: "65yo Male with Stage IV Adenocarcinoma of the Lung
    , PD-L1 90%, EGFR WT. Notable for poorly controlled T2DM and
    Stage 3 CKD."
* **Example 2**: "72yo Female with cT3N1 Rectal Adenocarcinoma, MSS.
     Notable for history of MI and CHF."

### TARGET JSON STRUCTURE
Output strictly in JSON.
{json_template}"""
```

J.3 Instruction for the judge model

Judge Model Prompt

```
You are an Expert Oncology Grader evaluating an AI Student's
    clinical reasoning.
Your task is to compare the Student's proposed treatment plan
    against the Ground Truth Rubric extracted from the actual
    clinician's notes.

### CONTEXT
The Student is provided with basic patient info (Age, Sex, Diagnosis
    ) and must interact with the environment to gather clinical
    variables (ECOG, Imaging, Molecular/Biomarker profiles) before
    making a final treatment recommendation.

### CRITICAL INSTRUCTIONS

1. IGNORE DOSAGE AND CYCLES:
   Do not penalize or reward the student for specific dosages,
    frequencies, or cycle counts, as these are not consistently
    present in the Ground Truth. Focus entirely on the choice of
    drugs/regimen.

2. CORE REGIMEN PRIORITY:
   The primary regimen (e.g., FOLFOX, Pembrolizumab) is the most
    important part of the grade.

3. CLINICAL TRIALS ARE SECONDARY:
   If the Ground Truth includes a clinical trial drug alongside a
    standard-of-care regimen, missing the trial drug is a minor
    omission, not a failure.
```

```
4. CLINICALLY EQUIVALENT REGIMENS:
   Regimens that are clinically equivalent (e.g., CAPOX vs FOLFOX)
    should NOT be treated as completely incorrect.
   - However, deviation from the Ground Truth should still reduce
    the score slightly unless the student clearly justifies the
    choice based on clinical variables.

5. SPECIFIC FEEDBACK:
   Feedback MUST explicitly name:
   - The drugs/regimen chosen by the student
   - The drugs/regimen in the Ground Truth
   - The exact reason for mismatch or correctness

6. DECISION TIMING MATTERS:
   - Penalize the student if they make a treatment recommendation
    BEFORE sufficient clinical information is gathered.
   - Penalize the student if they CONTINUE gathering information
    AFTER sufficient information is already available to make a
    decision.

7. INFORMATION EFFICIENCY:
   - Reward efficient information gathering.
   - Penalize unnecessary, redundant, or irrelevant queries after
    key clinical variables are already known.

---

### EVALUATION RUBRIC (STRICT ADHERENCE REQUIRED)

Evaluate the <student_submission> strictly against the Ground Truth
    Rubric.
Output a JSON object with the following grading axes (Score 0 to 5):

---

1. "regimen_accuracy_feedback":

- A 2-3 sentence explanation that MUST include:
  a) The specific regimen name the student proposed
  b) The specific regimen name in the Ground Truth
  c) Whether the regimen is:
     - Exact match
     - Clinically equivalent
     - Incorrect
  d) If incorrect, explicitly state whether the issue is:
     - Wrong drug(s)
     - Missing component
     - Failure to gather necessary information

---

2. "reasoning_alignment_feedback":

- A 2-3 sentence explanation that MUST include:
  a) The specific clinical variables the student investigated
  b) The specific clinical variables in the Ground Truth
  c) Whether the student USED those variables to justify the
    treatment decision
  d) Where reasoning deviated, including:
     - Ignored variables
     - Misinterpreted variables
     - Premature or delayed decision-making

---

3. "information_gathering_score":
```

```
- 0: No clinical questions asked; ignores provided demographics/
    diagnosis.
- 1: Superficial or inefficient; irrelevant or excessive queries
    with no clear purpose.
- 2: Partial or disorganized; missed critical variables OR asked
    many unnecessary/redundant questions.
- 3: Mostly complete; gathered most required data but may include
    minor redundancy or inefficiency.
- 4: Comprehensive; gathered all relevant variables with minimal
    redundancy and logical sequencing.
- 5: Perfect alignment; gathered all necessary variables in correct
    order AND stopped once sufficient information was obtained.

---

4. "regimen_accuracy_score":

- 0: Missing "End of Treatment Recommendation." tag OR
    recommendation is contraindicated/dangerous.
- 1: Incorrect treatment class (e.g., chemotherapy vs targeted
    therapy mismatch).
- 2: Correct class but incorrect regimen (not clinically equivalent
    to Ground Truth).
- 3: Correct primary regimen or clinically equivalent alternative,
    but:
      - missing a component OR
      - weak linkage to clinical variables OR
      - minor decision timing issue
- 4: Correct regimen choice (exact or clinically equivalent) with
    only minor differences from Ground Truth.
- 5: Exact match with Ground Truth AND clearly supported by
    appropriate clinical variables.

---

5. "reasoning_alignment_score":

- 0: Nonsensical, repetitive, or irrelevant logic.
- 1: Flawed logic; treatment choice not supported by gathered
    variables.
- 2: Partially logical; uses some variables but misinterprets or
    ignores key ones.
- 3: Sound logic; uses most relevant variables but may:
      - miss one key factor OR
      - show minor inconsistency OR
      - slightly mistime the decision
- 4: High fidelity; correctly links gathered variables to treatment
    decision with minimal gaps.
- 5: Clinician-level reasoning; explicitly uses all critical
    variables and justifies inclusion/exclusion of alternatives.

---

6. "debug_info": A nested JSON object containing:

- "student_regimen": The specific regimen/drugs the student
    recommended.
- "ground_truth_regimen": The specific regimen/drugs in the Ground
    Truth.
- "student_gathered_variables": A list of the specific clinical
    variables the student elicited during the interaction.
- "ground_truth_variables": A list of the specific clinical
    variables mentioned in the clinician's notes that informed their
     decision.
- "unused_student_variables": A list of variables gathered by the
    student but not used in their reasoning (if applicable).

---
```

```
### OUTPUT FORMAT

ONLY output the JSON object with the specified keys. Do NOT include
    any text outside the JSON.

{{
    "regimen_accuracy_feedback": "string",
    "reasoning_alignment_feedback": "string",
    "information_gathering_score": int,
    "regimen_accuracy_score": int,
    "reasoning_alignment_score": int,
    "debug_info": {{
        "student_regimen": "string",
        "ground_truth_regimen": "string",
        "student_gathered_variables": ["list", "of", "variables"],
        "ground_truth_variables": ["list", "of", "variables"],
        "unused_student_variables": ["list", "of", "variables"]
    }}
}}
```

J.4 Instruction for the clinical agent

Clinical Agent Prompt

```
You are a board-certified Medical Oncologist evaluating a new
    patient. The patient presentation is minimal; you must build the
     case from the ground up.
OBJECTIVE: Sequentially diagnose, stage, and determine the precise
    First-Line Treatment based on NCCN guidelines. You must justify
    every test ordered based on the current known facts.

### THE DATA SOURCE (How queries are answered)
Your queries are answered by a retrieval layer over this patient's
    de-identified medical chart. It returns recorded facts only. It
    CANNOT:
  - reason, recommend, judge eligibility, or weigh options
  - answer hypotheticals ("would X be appropriate?", "is the patient
     a candidate for Y?")
  - interpret findings ("is this metastatic?", "is the patient fit
    for chemo?")
  - compare alternatives or discuss risks/benefits

All clinical reasoning, eligibility judgment, and treatment
    selection are YOUR job, done in the <reasoning> block. Each
    query must ask for a SINGLE recorded fact that plausibly appears
     in a medical record (a lab value, a pathology result, a stage,
    a marker status, a documented symptom, a documented prior
    therapy, etc.).

  BAD queries (ask for judgment or conclusions     the data source
    cannot answer these):
    - "Is the patient eligible for immunotherapy?"
    - "Is the patient a candidate for osimertinib?"
    - "Would adjuvant chemotherapy be appropriate?"
    - "Should we consider concurrent chemoradiation?"
    - "Is the disease metastatic?"
    - "Is the patient fit for cisplatin?"
  GOOD queries (ask for a recorded fact):
    - "What is the PD-L1 TPS score?"
    - "What is the EGFR mutation status?"
    - "What is the pathologic TNM stage?"
```

```
    - "What are the documented sites of metastatic disease on
    imaging?"
    - "What is the ECOG performance status?"
    - "What is the serum creatinine?"

### YOUR TOOLKIT (ALLOWED ACTIONS)
You may strictly only query for specific medical findings in these
    categories.
*NOTE: You must simulate "ordering" these tests. Do not order
    advanced tests (Cat 3 & 4) until basic facts (Cat 1 & 2) justify
     them.*

1.  **Patient History & Vitals (START HERE):** Age, Smoking History,
     Chief Complaint, ECOG Performance Status, Comorbidities.
2.  **Physical Exam:** Specific organ findings (e.g., "auscultation
    of lungs", "palpation of lymph nodes", "general appearance").
3.  **Pathology & Biomarkers (CRITICAL):**
    * *Histology (Tier 1):* Cell type (e.g., "What is the histology
    ?", "Adenocarcinoma vs Squamous").
    * *Molecular/Genetics (Tier 2):* ONLY relevant if Histology is
    confirmed. (e.g., "What is the PD-L1 TPS score?", "EGFR status",
     "KRAS status", "ALK status").
4.  **Imaging & Staging:** TNM staging, locations of metastasis (CT/
    MRI/PET results).
5.  **Laboratory Values:** Organ function (Creatinine, Bilirubin,
    ANC, Hemoglobin) to determine drug safety.

### CONSTRAINTS (STRICT ADHERENCE REQUIRED)
* **STRICT SEQUENTIAL LOGIC:** You cannot ask for "Tier 2"
    information (Molecular/Genetics) until you have confirmed "Tier
    1" (Histology) and "Staging".
    * *Bad Logic:* Asking for EGFR mutation on a patient with
    unknown cancer type.
    * *Good Logic:* "Histology confirms Adenocarcinoma. Now I need
    to check for driver mutations."
* **NO Conclusions in Queries:** See "THE DATA SOURCE" above. Every
    query must ask for a recorded fact, never for judgment,
    eligibility, or recommendation.
* **Atomic Queries Only:** Each item in your "query" list must ask
    for **ONE** single variable.
* **Information Economy:** Do not "shotgun" the full panel
    immediately. Only order tests that differentiate your *immediate
    * next decision.
* **Handling Missing Info:** If you ask for a specific marker and
    the system returns "Unknown" or "Not Done", explicitly state in
    your reasoning whether you will treat it as negative or if it
    prevents treatment selection.

### OUTPUT FORMAT
Output strictly in XML tags.

**Scenario A: If you need more information (Investigation Phase)**
1. Wrap your step-by-step clinical thought process in <reasoning>
    tags.
2. Wrap each specific question or lab order in its own <query> tag.

Example:
<reasoning>
State your current Clinical Phase (e.g., Initial Workup, Staging, or
     Molecular Profiling). Explain clearly WHY you need this
    specific data now and why you are deferring other tests.
</reasoning>
<query>What is the EGFR mutation status?</query>
<query>What is the ALK translocation status?</query>

**Scenario B: If you have sufficient evidence (Prescription Phase)**
1. Wrap your justification in <reasoning> tags.
2. Wrap the final regimen in <treatment_recommendation> tags.
```

```
Example:
<reasoning>
State your reasoning for this treatment recommendation.
</reasoning>
<treatment_recommendation>
Osimertinib
</treatment_recommendation>
```