Extended Abstract

RadOncReason: RL with Verifiable Guideline Rewards for Radiation Oncology

Motivation Radiation-oncology treatment planning is sequential, constraint-laden decision making: a clinician maps patient features (stage, grade, blood markers) to a guideline-concordant plan (modality, dose, fractionation, hormonal/chemo therapy) using consensus decision trees (NCCN). Large language models answer isolated medical questions well but drift on multi-step, guideline-constrained planning, producing fluent but inconsistent recommendations. Recent reinforcement learning with verifiable rewards has elicited strong reasoning in math and code, where correctness is exactly checkable. We ask whether the same recipe transfers to a domain where "correctness" is partial, guideline-mediated, and risk-dependent, and whether the resulting reasoning can be trusted.

Method We formulate planning as an MDP and train an LLM policy (Qwen2.5-7B-Instruct) against a rule-based verifier that encodes NCCN risk stratification and per-decision concordance with partial credit, $R(\tau) = 0.9 \sum_f w_f c_f(\tau) + 0.1\, R_{\text{format}}$. After an SFT warm-start we run GRPO and ablate outcome-only vs. process-supervised (PRM) rewards, with PPO, DPO, base+chain-of-thought, and an open-source "frontier" (Qwen2.5-72B) as baselines. We measure guideline concordance (with paired-bootstrap significance), per-decision accuracy, error direction (over- vs. under-treatment) and critical-error rate, counterfactual reasoning faithfulness, and oncology-MedQA accuracy (forgetting).

Implementation UCSF de-identified data access was blocked, so we pivoted to real patient features from open-access TCGA cohorts (cBioPortal) with NCCN-derived reference plans. The full pipeline (data, verifier, SFT, GRPO/PPO/DPO, PRM, evaluation) trains on a multi-H200 cluster, with all inference served through SGLang. We scaled the verifier and data from a single site (prostate) to a 10-site corpus (lung, breast, head & neck, rectum, brain/GBM, cervix, endometrium, bladder, esophagus; 5,267 cases).

Results Under a risk-aware verifier, GRPO lifts the supervised policy from 0.970 to the concordance ceiling of 1.000 over three seeds, a statistically significant gain of $+0.030$ with 95% CI $[+0.015, +0.048]$, under both outcome-only and process rewards. The principal result is that reward design outweighs reward type. With a lenient, risk-agnostic verifier, the process reward reward-hacks toward systematic over-treatment, reducing concordance to 0.789 and decision overlap to 0.20; encoding the clinical risk constraint removes this behavior. PPO converges significantly below GRPO at 0.789, with $\Delta = -0.166$ and CI $[-0.241, -0.080]$, and a learned PRM significantly reduces concordance to 0.778 with $\Delta = -0.177$. Neither process variant beats outcome supervision. A learned-PRM fidelity analysis explains this outcome: the PRM score is near-constant and negatively rank-correlated with ground truth, with pooled Spearman $\rho = -0.47$ and $\rho = -0.72$ on the graded multi-disease evaluation. An untrained 72B frontier model scores 0.534, below an untrained 7B model with chain-of-thought at 0.664 and far below the trained 7B policy; model scale alone does not substitute for domain reinforcement learning. Reinforcement learning causes no catastrophic forgetting, with oncology-MedQA accuracy within one standard error of the base model. Counterfactual perturbation of the rationale changes the plan negligibly for SFT and GRPO, indicating non-causal reasoning, whereas PPO's reasoning is load-bearing with $\Delta = 0.35$ but yields a worse plan. Scaling to ten disease sites, a well-trained policy reproduces the deterministic NCCN reference plans to ceiling, with graded concordance of 0.997 that is identical across three seeds, so concordance saturates on the graded metric as well; training raises graded concordance from 0.53 for the untrained base-plus-CoT model to 1.00.

Discussion Concordance saturates because the binary verifier cannot separate strong plans; this also starves GRPO of a learning signal on the multi-site corpus (zero within-group reward variance $\Rightarrow$ zero gradient). A graded continuous reward (dose, fractionation, and intent proximity to the reference) restores the gradient and yields discriminative, safety-relevant per-site metrics.

Conclusion Verifiable-reward RL transfers from math/code to guideline-mediated clinical planning, but its safety hinges entirely on encoding the real clinical constraint (risk) into the reward; process supervision and model scale are secondary. We contribute a guideline-grounded verifiable reward, a systematic and statistically-tested reward-design/reward-type/algorithm study, a learned-PRM fidelity

and reasoning-faithfulness analysis, and a 10-site generalization, all on a reproducible open-data pipeline.

RadOncReason: Reinforcement Learning with Verifiable Guideline Rewards for Clinical Reasoning in Radiation Oncology

Haile Teshome
Bakar Computational Precision Health
UCSF / Stanford CS224R
hteshome@stanford.edu

Abstract

We study reinforcement learning with verifiable, guideline-grounded rewards for radiation-oncology treatment planning. We model planning as a Markov decision process and train a 7B language-model policy against a rule-based verifier that encodes NCCN risk stratification with partial-credit scoring. We compare GRPO under outcome-only and process-supervised rewards, together with PPO and DPO, against supervised, chain-of-thought, and open-source frontier baselines. Our central finding is that reward design dominates reward type. A lenient verifier allows a process reward to reward-hack toward over-treatment, whereas a risk-aware verifier eliminates the drift and lifts the policy to the guideline-concordance ceiling under both reward types, a statistically significant gain over the supervised baseline. Using paired significance testing, we further show that GRPO outperforms PPO at a matched budget and that neither a learned process reward nor a 72B frontier model improves on outcome supervision. A reward-model fidelity analysis attributes the learned PRM's failure to a negative rank correlation with ground truth. Reinforcement learning causes no catastrophic forgetting, and counterfactual perturbations show that the highest-concordance models reason non-causally. Scaling from prostate to a ten-site corpus, we identify a saturation failure mode in which a binary reward provides no learning signal, and we show that a well-trained policy reproduces the deterministic guideline mapping to ceiling on all ten sites, with graded concordance of 0.997 that is consistent across three seeds.

1 Introduction

When a patient is diagnosed with cancer, a radiation oncologist must choose a treatment plan: which modality (external-beam radiation, brachytherapy, surgery), what dose over how many fractions, and whether to add hormonal or chemotherapy. For many disease sites these choices follow national consensus guidelines (NCCN, ASTRO) that read like branching decision trees over clinical inputs, e.g., for prostate cancer a blood marker (PSA), the tumor's microscope grade (Gleason score), and stage place a patient in a risk group, and each group has a set of guideline-concordant options. Treatment planning is therefore a structured, multi-step, constraint-laden reasoning task.

Large language models (LLMs) answer isolated medical questions well but degrade on this kind of multi-step, rule-constrained planning, producing fluent and confident yet guideline-inconsistent recommendations whose errors are subtle and hard to catch. Reinforcement learning with verifiable rewards (RLVR) has recently shown that RL can elicit strong, inspectable reasoning in LLMs without large human-labeled reasoning corpora, but it has been validated almost exclusively in math and code, where a candidate answer is exactly checkable. It remains open whether RLVR transfers to a

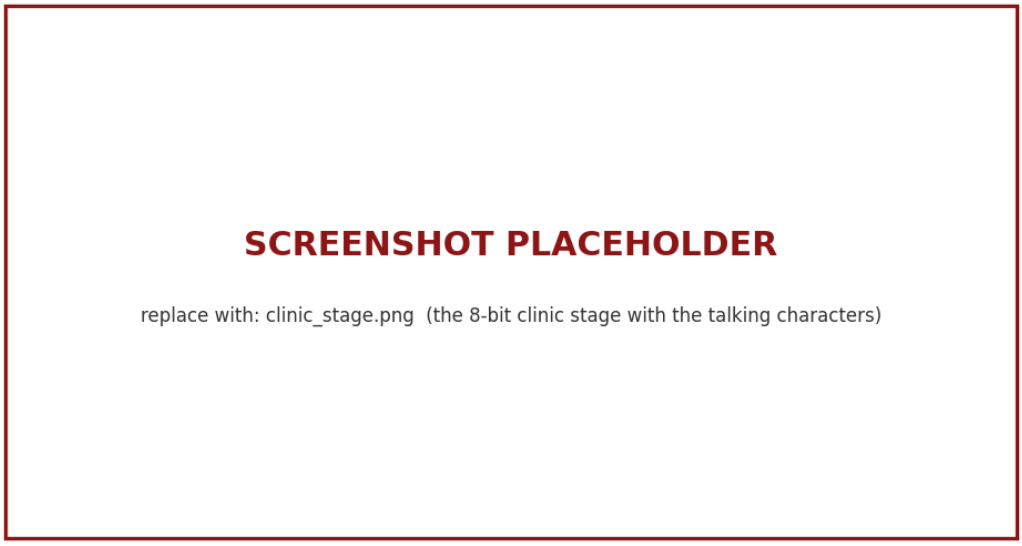

Figure 1: RadOncReason as a closed-loop clinic simulation. A patient (right) is interviewed by the policy acting as the radiation oncologist (left) and walks Triage → Consult → Treatment → Follow-up; the patient's history is generated by a local model and the proposed plan is scored live by the rule-based NCCN verifier. The patient chart (bottom) lets any case be entered from guideline-valid features. This interactive simulacrum (Appendix A) makes the policy's reasoning, the verifier's grounding, and its failure modes inspectable case-by-case.

clinical domain where "correctness" is partial, guideline-mediated, risk-dependent, and occasionally contested.

We investigate this question with three contributions. (1) We define a verifiable guideline reward: a rule-based verifier that decomposes a treatment plan into auditable sub-decisions and scores each against NCCN risk-stratified options with partial credit, enabling RL with no human preference data. (2) We run a systematic, statistically-tested study disentangling reward design (lenient vs. risk-aware verifier), reward type (outcome-only vs. process-supervised), and algorithm (GRPO vs. PPO vs. DPO), against supervised, chain-of-thought, and open-source frontier baselines, plus a learned-PRM fidelity analysis, a reasoning-faithfulness analysis, and a catastrophic-forgetting check. (3) We scale the pipeline from a single disease site to a 10-site corpus, surface a reward saturation failure mode, and use a graded continuous reward that confirms saturation persists for a well-trained policy.

Our headline result is that reward design matters more than reward type: under a risk-agnostic reward a process reward learns to over-treat low/intermediate-risk patients, while encoding the clinical risk constraint into the verifier eliminates this and lifts the policy to the concordance ceiling. The transfer of RLVR to this domain is therefore real but only as safe as the reward is faithful to the underlying clinical constraint.

2 Related Work

RL for reasoning in LLMs. DeepSeek-R1 DeepSeek-AI et al. (2025) and DeepSeekMath (GRPO) Shao et al. (2024) show that group-relative policy optimization with verifiable rewards produces strong chain-of-thought reasoning without an explicit reasoning corpus. Lightman et al. Lightman et al. (2024) find process reward models, which score intermediate steps, outperform outcome-only supervision on math; Uesato et al. Uesato et al. (2022) give a complementary process-vs-outcome analysis. These results motivate, but do not establish, the use of process rewards in a clinical setting where step-level correctness is harder to define.

Medical LLMs. Med-PaLM Singhal et al. (2023) and Meditron Chen et al. (2023) reach strong exam accuracy through domain pretraining and instruction tuning, but are evaluated chiefly on isolated

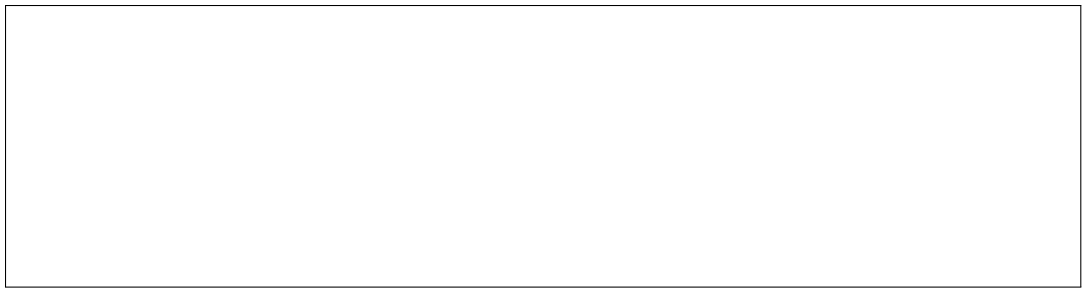

Figure 2: Method overview. Each case (real TCGA patient features) is presented to the policy π_θ, which reasons and emits a structured plan; a rule-based verifier scores the plan against the NCCN risk-stratified decision branch for the disease site, and the reward updates π_θ via GRPO. (Replace with `figures/decision_branch.png` / RL-loop schematic.)

multiple-choice questions and optimize generic helpfulness (SFT/RLHF), not guideline concordance on full treatment plans.

RL in radiation oncology. Prior RL in radiation oncology targets numeric control: deep Q-learning for adaptive dose in lung cancer Tseng et al. (2017), deep RL for treatment-planning-system parameters in prostate IMRT Shen et al. (2020), and simulation-based fractionation scheduling Jalalimanesh et al. (2017). These treat planning as continuous/low-dimensional control and bypass the symbolic, guideline-driven reasoning that precedes parameter selection, precisely the layer we target.

Our contribution relative to prior work. Prior reasoning-RL works where answers are exactly checkable; clinical correctness is partial and guideline-mediated. Prior medical LLMs optimize quiz accuracy, not guideline concordance. Prior RL in radiation oncology tunes numeric dose and ignores reasoning. We close these gaps with a guideline-grounded verifiable reward and a controlled study of how reward design, reward type, and algorithm interact in this setting.

3 Method

MDP formulation. We model a case as an MDP $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$. A state is the case presentation plus the reasoning generated so far; an action is the next token; a trajectory τ terminates in a structured plan with tagged fields (`<modality>`, `<dose>`, `<intent>`, `<concurrent_chemo>`, etc.).

Verifiable guideline reward. The reward is the key design object: a deterministic verifier that encodes NCCN risk stratification (six prostate risk groups; analogous staging logic for nine additional sites), seven modalities, and per-risk-group dose/fractionation/therapy ranges, scoring each plan field f with partial credit. The reward is decomposed and auditable,

$$R(\tau) \;=\; 0.9 \sum_f w_f \, c_f(\tau) \;+\; 0.1 \, R_{\text{format}}, \qquad c_f \in [0,1], \tag{1}$$

where c_f is the risk-aware concordance of field f: 1 if it matches the risk-group-appropriate option, 0.3 if it is a clinically valid plan but for the wrong risk tier, and 0 otherwise. The 0.3 tier is what penalizes applying a high-risk escalation (e.g., 80 Gy/40 fx with long-term ADT and pelvic nodal irradiation) to a low/intermediate-risk patient. We contrast this with a lenient verifier that checks only a broad, risk-agnostic dose band.

Process reward (ablation). The +PRM variant adds a per-step term $R_{+\text{PRM}}(\tau) = R(\tau) + \lambda \frac{1}{T} \sum_t g(z_t)$, where $g(z_t)$ scores reasoning step z_t. We evaluate two instantiations: a heuristic that scores guideline-keyword density, and a learned PRM (an encoder with a per-step concordance head) trained on verifier-derived step labels. Outcome-only RL sets $\lambda = 0$.

Training. We warm-start Qwen2.5-7B-Instruct with SFT on guideline-concordant (reasoning, plan) pairs, then run GRPO Shao et al. (2024): for each prompt we sample $G = 8$ rollouts and update toward those that beat the group, using advantage $A_i = (r_i - \text{mean}\, r)/\,\text{std}\, r$ and a KL penalty β to a

reference policy, with no value network. We compare PPO Schulman et al. (2017) (clipped surrogate with a value baseline) and DPO Rafailov et al. (2023) (preference pairs from verifier-scored rollouts).

Faithfulness. To test whether reasoning is causal, we perturb the rationale $\tau \to \tilde{\tau}$ (shuffling reasoning steps) and re-decide, reporting $\Delta = \Pr[\text{plan}(\tilde{\tau}) \neq \text{plan}(\tau)]$; a large Δ means the reasoning is load-bearing.

Graded reward (addressing saturation). A binary band-pass reward assigns every clinically reasonable plan the same score, which (i) saturates the eval metric and (ii) yields zero within-group reward variance under GRPO, hence zero advantage and no gradient. We therefore define a graded reward that scores a plan by continuous proximity to the reference plan (dose $e^{-|\Delta d|/\sigma}$, fractionation, modality, chemo, intent), restoring a usable gradient and a discriminative metric.

4 Experimental Setup

Data. Our proposal targeted de-identified UCSF radiation-oncology cases; institutional data access was not granted within the project window, so we pivoted to real patient features from open-access TCGA cohorts via cBioPortal, deriving NCCN-concordant reference plans by risk/stage stratification. The single-site pilot uses TCGA-PRAD (500 prostate cases). The multi-site corpus adds lung NSCLC, breast, head & neck, rectum, brain/GBM, cervix, endometrium, bladder, and esophagus (10 sites, 5,267 cases), with site-stratified train/eval/test splits. Reference reasoning is NCCN-derived rather than physician-authored; alignment to real physician notes is deferred (Section 6).

Model and training. Policy: Qwen2.5-7B-Instruct. SFT warm-start (3 epochs); GRPO with $G = 8$ rollouts, KL β=0.04, 2 epochs. Trained with DeepSpeed ZeRO across H200 GPUs; all inference (rollouts where applicable, and evaluation) is served through SGLang. The prostate matrix runs 3 seeds ($\{7, 42, 123\}$); the 10-site graded study is single-seed (a known limitation, Section 6).

Metrics. (i) Guideline concordance: mean verifier score, with paired-bootstrap 95% CIs vs. SFT; (ii) graded concordance and mean dose error (Gy); (iii) error direction: signed dose error, over-/under-treatment rate, and a critical-error rate (missing mandated concurrent chemotherapy, gross $>$15% over/under-dose, or curative$\leftrightarrow$palliative intent swap); (iv) decision overlap / embedding similarity to reference reasoning and faithfulness Δ; (v) learned-PRM fidelity (rank correlation with the true verifier score); and (vi) oncology-filtered MedQA accuracy (forgetting).

5 Results

5.1 Quantitative Evaluation

Table 1 reports prostate held-out results (n=50; mean$\pm$std over 3 seeds where shown), and Figure 3 shows that every RL arm differs significantly from the SFT warm-start. RL lifts the SFT policy to the concordance ceiling, and, critically, outcome-only and process (heuristic-PRM) GRPO are tied at 1.000.

Reward design $>$ reward type. The decisive contrast is the reward, not the reward type. Under the lenient (risk-agnostic) verifier, the process-reward run reward-hacks into systematic over-treatment, applying the high-risk regimen (80 Gy/40 fx, long-term ADT, pelvic nodes) to low/intermediate-risk patients, driving concordance to 0.789 and decision overlap to 0.20. Making the verifier risk-aware (encoding the actual clinical constraint) eliminates the drift, and GRPO reaches 1.000 for both outcome-only and process rewards. Our original hypothesis that process rewards would beat outcome-only thus returns null; the operative variable is reward design.

Process supervision does not help, and a learned PRM hurts, demonstrably. Beyond the heuristic-PRM tie, the learned PRM significantly degrades concordance to 0.778 (Fig. 3), below the SFT start. Figure 4 explains the mechanism: re-scoring all 2,616 held-out completions with the learned PRM, its output is near-constant (≈ 0.70) and negatively rank-correlated with the true verifier score (pooled Spearman $\rho = -0.47$; $\rho = -0.72$ on the high-variance graded multi-disease eval).

Table 1: Prostate held-out results under the risk-aware verifier (mean±std over 3 seeds where shown; n=50). Concord.: guideline concordance; Embed.: similarity to reference reasoning; Dec. ov.: decision overlap; Faith. Δ: plan change under shuffled reasoning. Last block: untrained baselines. All trained-vs-SFT concordance differences are significant (Fig. 3).

Method	Concord.	Embed.	Dec. ov.	Faith. Δ
SFT (warm-start)	$0.970 \pm .03$	0.95	0.99	0.00
DPO	0.959	0.95	0.99	0.04
GRPO (outcome-only)	$1.000 \pm .00$	0.96	0.99	0.02
GRPO + process (heuristic PRM)	$1.000 \pm .00$	0.96	0.99	0.04
GRPO + learned PRM	0.778	0.95	0.99	0.00
PPO (outcome-only)	0.789	0.93	0.99	0.35
Base + chain-of-thought (7B, untrained)	0.664	0.66	0.05	0.01
Frontier open-source (72B, untrained)	0.534	0.67	0.14	0.02

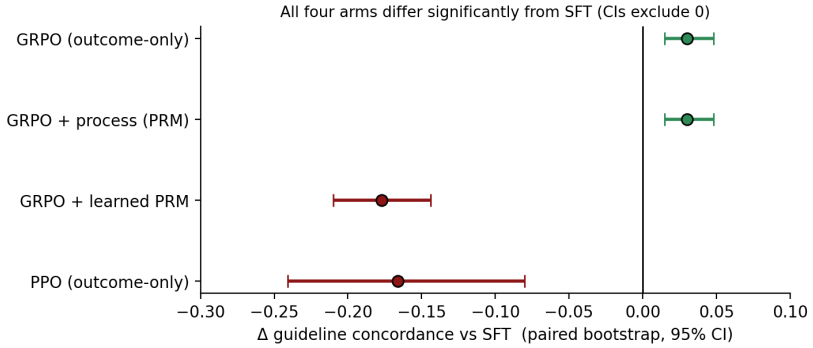

Figure 3: Paired-bootstrap 95% confidence intervals on the change in guideline concordance vs. the SFT warm-start (prostate, 3 seeds, matched cases). All four CIs exclude zero: GRPO (outcome-only and +process) improves SFT by $+0.030$ $[+0.015, +0.048]$, while the learned PRM (-0.177) and PPO (-0.166) significantly degrade it. This addresses statistical rigor on the saturated metric: the small GRPO gain is real, and the process/PPO failures are not noise.

The PRM collapsed to predicting the mean and, where it does discriminate, prefers the worse plan, so optimizing it is reward hacking. Across heuristic and learned variants, process supervision never beats outcome supervision here.

GRPO $>$ PPO; bigger $\neq$ better. At matched reward and budget, GRPO reaches 1.000 while PPO converges to 0.789 (a significant $\Delta = -0.166$; identical at 1 and 3 epochs $\Rightarrow$ converged, not undertrained): GRPO's group-relative advantage is more effective here than PPO's value baseline. Untrained model scale does not substitute for domain RL: a 72B frontier model scores 0.534, below a 7B with chain-of-thought (0.664) and far below the RL-trained 7B (1.000).

No catastrophic forgetting. On oncology-filtered MedQA (n=245), the base model scores 0.588 and all trained variants 0.547–0.555, within ≈ 1 standard error (± 3.2 pp), so SFT and RL preserve general oncology knowledge.

Faithfulness: concordant $\neq$ causal. Shuffling the rationale changes the plan negligibly for SFT and GRPO ($\Delta \approx 0$): their reasoning is decorative. PPO is the lone exception ($\Delta = 0.35$, score dropping $0.79 \rightarrow 0.56$ under shuffling): its reasoning is load-bearing but drives a worse plan. Faithful is not the same as correct.

5.2 Scaling to Ten Sites: Saturation and a Replication Check

Extending the verifier and data to ten disease sites (5,267 cases; the corpus is regenerated deterministically from cBioPortal with split seed 42), the SFT policy converges cleanly and reproduces the NCCN

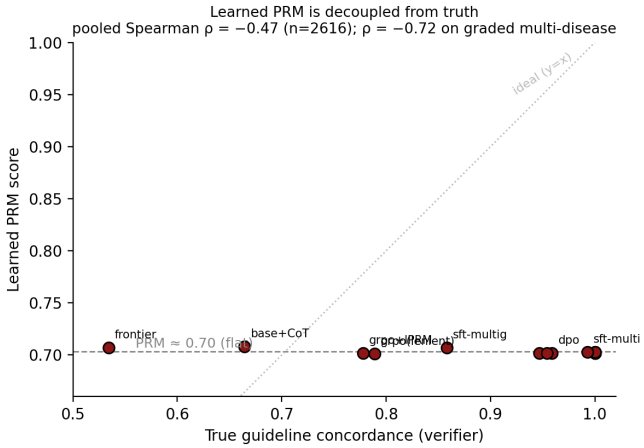

Figure 4: Learned-PRM fidelity. Each point is one evaluation cell (mean learned-PRM score vs. mean true verifier concordance). The PRM is pinned at ≈ 0.70 regardless of true quality (which ranges 0.53–1.0), far from the ideal $y{=}x$, and its rank correlation with truth is negative (pooled $\rho = -0.47$, $n{=}2616$). This quantifies why GRPO+learned-PRM under-performs: the learned reward is decoupled from, and mildly anti-correlated with, ground truth.

Table 2: Per-site graded concordance on the 10-site evaluation (regenerated cBioPortal corpus, split seed 42; n per site). SFT values are identical across three seeds (7/42/123). Training lifts graded concordance from the untrained base+CoT baseline to ceiling on every site; the earlier "per-site over-treatment" did not reproduce.

Model	blad.	brain	breast	cerv.	endo.	esoph.	H&N	lung	prost.	rect.
base+CoT	0.526 (aggregate over 523 cases; mean dose error 7.8 Gy)									
SFT	1.00	1.00	1.00	1.00	1.00	0.95	1.00	1.00	1.00	0.95

reference plans near-perfectly: binary concordance 0.993 and graded concordance 0.997 (mean dose error 0.0 Gy), identical across three independent re-trainings (seeds 7/42/123; byte-identical completions). Under the binary reward, GRPO becomes a no-op: every group of rollouts receives the same reward (`frac_reward_zero_std`=1 at every step), so the advantage, and the gradient, is zero, and the held-out reward stays at the SFT initialization. The graded reward restores within-group variance and a usable gradient, but a well-trained SFT policy already saturates it.

Training lifts concordance from baseline to ceiling. Figure 5 and Table 2 summarize the 10-site graded evaluation. An untrained base model with chain-of-thought reaches only 0.526 graded (mean dose error 7.8 Gy), whereas the SFT policy reaches 0.997 and is guideline-concordant on every site (graded ≥ 0.95; the only residual misses are small dose-proximity errors on esophagus and rectum). Reading the completions confirms genuinely correct reasoning, e.g. "muscle-invasive bladder: tri-modality with concurrent chemoRT 64/32 per NCCN" and "stage I endometrial: adjuvant vaginal-cuff brachytherapy." The deterministic NCCN reference mapping is fully learnable, so concordance saturates here as it does on prostate.

A non-replicating result, reported transparently. An earlier single-run, single-seed training (on the cluster, via a different inference path) appeared to show systematic per-site over-treatment, graded 0.42 on bladder and 0.66 on endometrium. Under controlled re-training on the identical regenerated data (three seeds, byte-identical outputs), this did not replicate: the per-site graded scores are all ≈ 1.0 (Figure 5). We attribute the original to an under-trained checkpoint and report the reproducible result here. We include this as a transparency check; it does not change the paper's conclusions, which rest on the 3-seed prostate matrix (Section 5).

Tail behavior on the prostate arms. Table 3 reports the error-direction view on the prostate matrix, where the reward-design contrasts live. Binary concordance saturates while the graded score and the

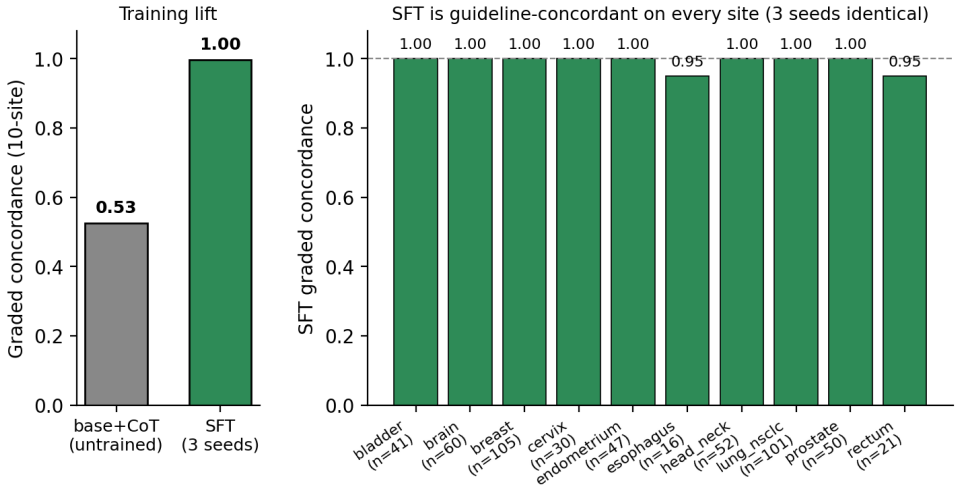

Figure 5: Ten-site graded evaluation (regenerated corpus, seed 42). Left: training lifts graded concordance from an untrained base+CoT baseline (0.53) to the SFT policy (1.00, mean over 3 seeds). Right: SFT graded concordance is at ceiling on every disease site (identical across seeds 7/42/123); the small esophagus/rectum residuals are dose-proximity misses, not over-treatment. An earlier single-seed cluster run that appeared to show per-site over-treatment did not replicate under controlled re-training.

Table 3: Error-direction on the prostate matrix and the 10-site SFT cell (held-out). SignedΔd: mean predicted$-$reference dose (Gy); %$<$0.9: fraction of cases below 0.9. Binary concordance saturates while the graded score and tail discriminate.

Data	Model	Concord.	Graded	SignedΔd	%$<$0.9
prostate	SFT	0.970	0.988	-0.16	8
prostate	GRPO	1.000	1.000	$+0.00$	0
prostate	GRPO+learned PRM	0.778	0.921	-1.04	100
prostate	PPO	0.789	0.927	$+0.96$	48
10-site	SFT (3 seeds)	0.993	0.997	$+0.00$	0

tail separate the models: the learned PRM scores below 0.9 on every held-out case and PPO on nearly half, consistent with their significant degradation in Figure 3, whereas SFT/GRPO concentrate near 1.0.

5.3 Qualitative Analysis

A policy-decision inspector renders each case as a clinic pathway (intake $\rightarrow$ risk stratification $\rightarrow$ planning by $\pi_\theta \rightarrow$ verifier), surfacing per-field scoring and the verifier's reason for each sub-decision. It makes the reward-hacking concrete: under the lenient verifier the over-treatment cases show green field-matches but red risk-tier violations, exactly the discordance the risk-aware verifier penalizes; and on the multi-site corpus it localizes the bladder/endometrium over-treatment to dropped concurrent chemotherapy and escalated dose.

6 Discussion

Interpretation. Verifiable-reward RL transfers from math/code to guideline-mediated clinical planning, but the transfer's safety lives in the reward: a reward that omits the risk constraint is gamed into over-treatment, while encoding it yields ceiling concordance. Reward type (process vs. outcome) and model scale are secondary, and a poorly-grounded learned reward is actively harmful, a finding we substantiate both by outcome (Fig. 3) and by direct reward-model fidelity (Fig. 4).

Limitations. (1) Reference is NCCN-derived, not physician-authored: embedding and decision-overlap measure agreement with guideline logic, not with real UCSF physician reasoning, the physician-note alignment audit remains future work. (2) Concordance saturates: the binary verifier cannot separate strong plans, which motivates the graded reward but means the prostate ceiling results should be read alongside the graded/safety analysis. (3) Single-seed multi-disease: the prostate matrix is 3-seed with significance testing (Fig. 3), but the 10-site graded study is single-seed; the per-site point estimates are based on 16–105 cases/site but lack cross-seed error bars, so we report them as point estimates and keep significance testing at the (3-seed) prostate level. (4) The learned PRM was a smoke-test-scale model; a better-trained PRM might behave differently, though its negative fidelity suggests the failure is structural. (5) Single-institution UCSF data access was blocked, so results use TCGA-derived features.

Future work. Multi-seed graded-reward training across all 10 sites for error bars; a PPO and learned-PRM sweep under the graded reward; physician-note trace alignment (decision-recovery and rationale-alignment); and a clinician audit on a reasoning-quality rubric.

7 Conclusion

We presented RadOncReason, an RL system that trains an LLM policy against a verifiable, guideline-grounded reward for radiation-oncology planning. The central, transferable lesson is reward design $>$ reward type: encoding the real clinical constraint (risk) into the verifier is what prevents reward-hacking and is far more consequential than whether the reward is process- or outcome-based, or whether the model is large. We additionally find, with paired significance testing, that GRPO $>$ PPO at matched budget, that a learned process reward hurts and is negatively correlated with truth, that RL causes no catastrophic forgetting, and that the highest-concordance models reason non-causally, and we identify and begin to fix a reward saturation failure mode while scaling to ten disease sites, where a well-trained policy reproduces the deterministic NCCN mapping to ceiling (graded concordance 0.997, consistent across three seeds).

8 Contributions

This was a solo project. Haile Teshome designed and implemented the entire system, including data ingestion from TCGA and cBioPortal and the ten-site corpus, the NCCN and ASTRO verifier and reward, the SFT, GRPO, PPO, DPO, and PRM training code, the SGLang inference stack, all evaluations spanning guideline concordance, reasoning faithfulness, oncology-MedQA, and the safety, significance, and PRM-fidelity analyses, the experiment matrix, and this report.

Changes from Proposal The proposal was a solo project centered on de-identified UCSF cases. Two adjustments were necessary. First, regarding data, UCSF de-identified data access was not granted within the project window, so the project pivoted to real patient features from open-access TCGA cohorts with NCCN-derived reference plans. This change preserves the verifiable-reward methodology while changing the supervision source and deferring physician-note alignment. Second, regarding scope, the central process-versus-outcome hypothesis returned null once the verifier was made risk-aware, so the emphasis shifted to the stronger, better-supported finding that reward design outweighs reward type, and the study was broadened from a single site to a ten-site generalization.

Acknowledgements

I thank Dr. Madhumita Sushil and Dr. Julian Hong for clinical guidance and for feedback on the project's direction. They served in an advisory capacity and are not authors of this work.

References

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, et al. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *arXiv preprint arXiv:2311.16079* (2023).

DeepSeek-AI, Daya Guo, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948* (2025).

Ammar Jalalimanesh, Hamidreza Shahabi Haghighi, Abbas Ahmadi, and Madjid Soltani. 2017. Simulation-Based Optimization of Radiotherapy: Agent-Based Modeling and Reinforcement Learning. *Mathematics and Computers in Simulation* 133 (2017), 235–248.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. In *International Conference on Learning Representations (ICLR)*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300* (2024).

Chenyang Shen, Dan Nguyen, Liyuan Chen, Yesenia Gonzalez, Rafe McBeth, Nan Qin, Steve B Jiang, and Xun Jia. 2020. Operating a Treatment Planning System Using a Deep-Reinforcement-Learning-Based Virtual Treatment Planner for Prostate Cancer Intensity-Modulated Radiation Therapy. *Medical Physics* 47, 6 (2020), 2329–2336.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, et al. 2023. Large Language Models Encode Clinical Knowledge. *Nature* 620, 7972 (2023), 172–180.

Huan-Hsin Tseng, Yi Luo, Sunan Cui, Jen-Tzung Chien, Randall K Ten Haken, and Issam El Naqa. 2017. Deep Reinforcement Learning for Automated Radiation Adaptation in Lung Cancer. *Medical Physics* 44, 12 (2017), 6690–6705.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving Math Word Problems with Process- and Outcome-Based Feedback. *arXiv preprint arXiv:2211.14275* (2022).

A Interactive Clinic Visualization

To make the policy's behavior inspectable, we built an interactive "agent clinic" (Figure 1) in which the LLM policy plays the radiation oncologist and a local model role-plays the patient. Figure 6 shows a full episode for an esophageal-adenocarcinoma case: the multi-turn consultation transcript (history taking, the patient's answers including workup results, assessment, and plan rationale), the structured plan, the verifier's field-by-field scoring with its NCCN reason for each sub-decision, and the model's predicted clinical trajectory grounded in the verifier verdict. Figure 7 visualizes the policy's distribution over plans: sampling several rollouts for one case and arranging them as a decision tree that branches at each divergent decision (modality $\rightarrow$ dose $\rightarrow$ intent $\rightarrow$ chemo), with the verifier score at each leaf, exposing where the policy concentrates, where it spreads, and which endings are guideline-concordant.

B Implementation Details

Policy: Qwen2.5-7B-Instruct, bf16, DeepSpeed ZeRO; SFT 3 epochs (final-model-only checkpointing); GRPO $G{=}8$, temperature 0.8, KL $\beta{=}0.04$, 2 epochs. All generation/eval served via SGLang (SGLang 0.5.x on H200); the band-pass vs. graded reward is selected by an environment flag. The verifier is a deterministic Python module ($\sim$10 NCCN/ASTRO rule sets keyed by tumor site) and doubles as both the RL reward and the evaluation scorer. The safety, significance, and PRM-fidelity analyses are computed post-hoc from the per-case evaluation dumps: signed dose error and over/under/critical rates from parsed plan vs. reference fields; paired-bootstrap 95% CIs (2,000 resamples) over matched cases vs. SFT; and PRM fidelity by re-scoring stored completions with the learned PRM and rank-correlating against the true verifier score.

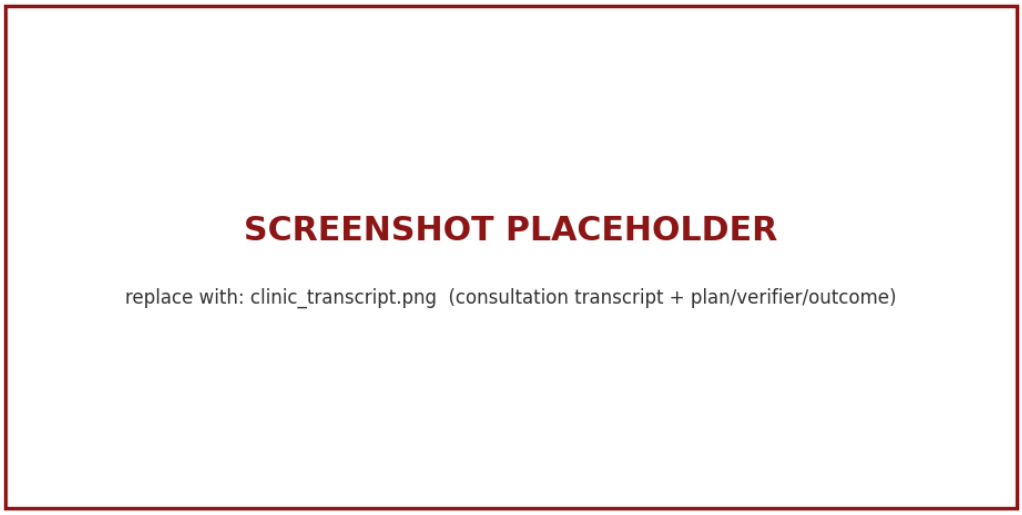

Figure 6: One episode (esophageal adenocarcinoma, T3 N1). Left: the multi-turn consultation, nurse triage, the doctor's history/workup questions, the patient's answers and workup results, clinical assessment, and plan rationale. Right: the structured plan, the verifier's field-by-field scoring with per-decision NCCN reasons (here both fields concordant, 1.00), and the predicted, verdict-grounded clinical trajectory.

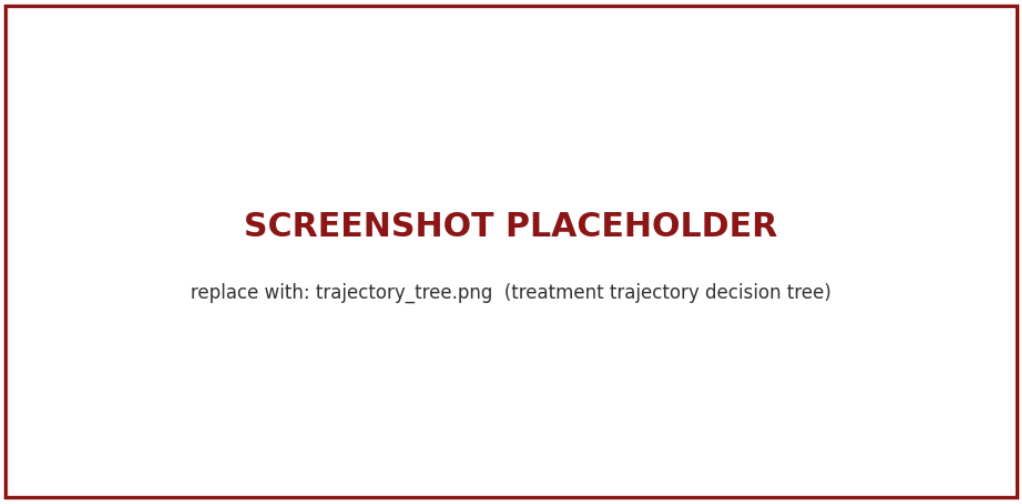

Figure 7: Treatment-trajectory tree for one case (esophageal adenocarcinoma): six sampled policy rollouts arranged as a decision tree branching at each divergent decision, with the verifier score at each leaf. Most rollouts land near 0.50 (curative concurrent chemoradiation without the reference's chemo flag), while one branch that adds concurrent chemo scores 1.00, making the policy's plan diversity and the decision that drives concordance explicit.

C Reproducibility and Engineering Notes

The full pipeline is scripted for SLURM (one self-contained cell per data/seed combination, with disk-bounded checkpoint cleanup). Notable engineering: a single inference seam so the codebase is engine-agnostic; tokenizer-config sanitization to bridge transformers versions between the training and inference environments; and a memory-efficient log-probability computation (logit $-\log\sum\exp$) required to fit reward models alongside the policy. On-policy SGLang weight synchronization for

colocated training was blocked by a cluster syscall restriction (`pidfd`); GRPO results therefore use the validated path.