LATENT DECISION FLOW REWARDS FOR CLINICAL GUIDELINE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical practice guidelines encode evidence-based recommendations as conditional decision flows, which makes them a natural source of process supervision for training language models. Using this supervision is difficult for two reasons. A decision flow consistent with a guideline does not by itself determine the correct answer for a given patient, which we refer to as the *path–answer gap*, and several valid decision flows can support the same clinical decision, which we refer to as *path multiplicity*. Under these conditions, answer rewards provide no signal about the decision process, supervised training on annotated flows constrains the model to a single trajectory and reduces answer accuracy, and process rewards computed against one reference flow penalize alternative valid reasoning. We introduce GUIDER, a joint controller–policy optimization approach that learns guideline-adherent clinical reasoning while accounting for these challenges. The controller learns to generate plausible decision flows from guideline-derived supervision and provides process-level supervision to the policy, while the policy learns to generate clinical reasoning and final decisions. By aggregating process rewards over multiple controller-generated decision flows, GUIDER accounts for path multiplicity, and by gating process rewards on answer correctness, it couples guideline adherence with correct final answers. Joint optimization allows process supervision to evolve with the policy instead of relying on a fixed reference trajectory. The policy is never conditioned on an oracle decision path, so GUIDER requires no access to guidelines at inference time. On MedGUIDE, GUIDER with Qwen3-14B achieves 82.5 weighted accuracy and 73.5 guideline adherence, improving over the strongest clinical reinforcement learning baseline by 3.3 and 3.1 points, respectively. GUIDER also improves over CRPO, MRPO, and FaithMed on the out-of-domain benchmarks AMEGA-LLM, RealMedQA, and CancerGUIDE with Qwen3-14B and Gemma4-12B backbones.

1 INTRODUCTION

Training large language models for clinical reasoning typically relies on rewards for final answer correctness or on preferences over complete responses (Jung et al., 2026). This supervision is sparse and does not identify which intermediate decision caused an error. In clinical decision support, this limitation is important because a recommendation depends on a sequence of conditional judgments about findings, contraindications, disease severity, and escalation thresholds. A model can also reach the correct answer through incomplete or poorly grounded reasoning, with a rationale that omits essential evidence or deviates from clinical guidelines (**?**Jiang et al., 2026; Farahani et al., 2026; Chen et al., 2026). These failures motivate training that rewards correct answers alongside decision pathways consistent with clinical guidelines.

Clinical practice guidelines offer a principled source of process supervision because they encode evidence-based recommendations as conditional decision structures instead of answer labels (Li et al., 2025; Unell et al., 2025; Fast et al., 2024; Kell et al., 2024). We refer to a sequence of clinical conditions and actions that leads toward an outcome as a *decision flow*. Recent work compiles guidelines into executable decision logic for training (Shen et al., 2026) and inference (Cao et al., 2026), and MedGuideX, for example, supervises each case with the single path executed by a deterministic decision tree. The decision flow underlying an answer, however, is *latent*, because the patient case and its answer do not uniquely determine it. As Figure 1 illustrates, different valid

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

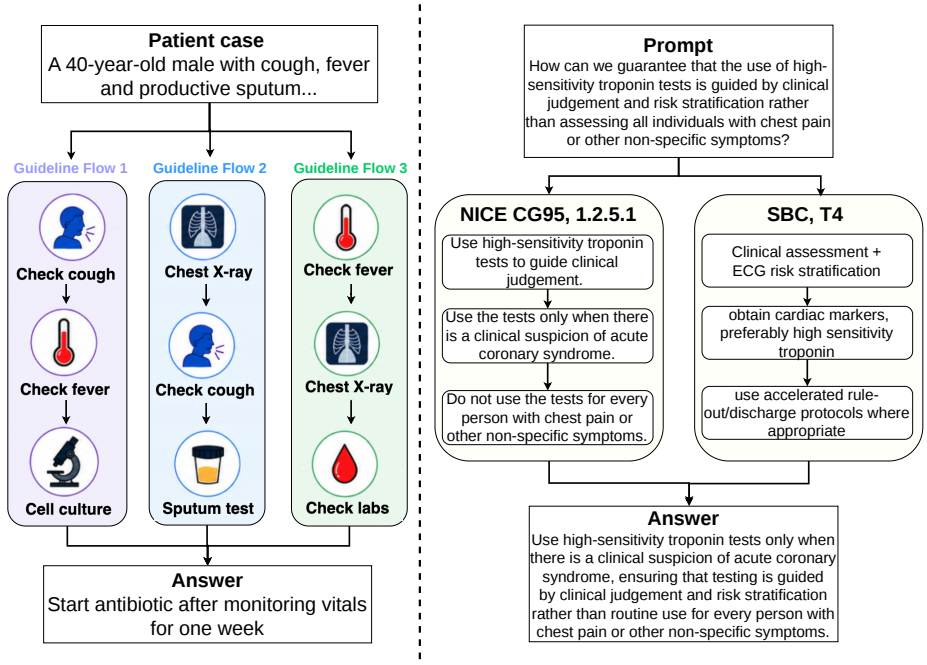


Figure 1: **Guideline decision flows illustrating the *path–answer gap* and *path multiplicity* in clinical reasoning. Left:** A hypothetical example illustrating multiple valid decision flows leading to the same clinical answer. **Right:** A real clinical example illustrating both phenomena: distinct decision flows from NICE and SBC guidelines support the same clinical answer, while neither flow directly specifies the complete answer. The respective decision flows are derived from NICE CG95 (Sec. 1.2.5.1–1.2.5.7) (National Institute for Health and Care Excellence, 2016) and the Brazilian Society of Cardiology (SBC, Table 4) (Feitosa-Filho et al., 2015).

guidelines can lead to the same clinical outcome through different decision flows. Two clinicians can likewise reach the same decision through different valid paths, depending on their experience and interpretation of the evidence, and the final decision does not reveal which path either of them followed. We refer to this property as *path multiplicity*. Under path multiplicity, supervision against a single reference flow is brittle, because it penalizes alternative reasoning that is also consistent with guidelines. Effective process supervision should therefore account for multiple plausible decision flows instead of requiring the model to reproduce one trajectory.

Reinforcement learning with verifiable rewards (RLVR) (Shao et al., 2024; DeepSeek-AI, 2025) has recently emerged as a promising approach for clinical reasoning, using verifiable outcome rewards to improve reasoning beyond supervised fine-tuning (Afolabi et al., 2026; Jiang et al., 2026; Jung et al., 2026). However, existing approaches largely optimize answer correctness or sequence-level reasoning quality, providing limited supervision over whether intermediate clinical decisions follow established guidelines. Clinical decision flows offer a natural source of process supervision by specifying relevant patient conditions, decision criteria, and recommended actions. Yet guideline-based process supervision introduces a key challenge: a valid decision flow does not necessarily determine the correct answer to a patient-specific query. Conversely, the same clinical outcome may be supported by multiple valid decision flows. We refer to this mismatch between process validity and outcome correctness as the *path–answer gap*, motivating an objective that models guideline adherence and answer correctness as complementary signals rather than treating them as aligned.

Present work. GUIDER (**Gu**ideline-**I**nformed **De**cision Flow Rewards for Clinical **R**easoning) is a reinforcement learning framework that incorporates guideline adherence into the reasoning process of LLMs. To address path multiplicity, we treat clinical decision flows as latent variables and represent guideline reasoning using multiple plausible trajectories rather than a single oracle path. A controller trained on guideline-derived reference flows generates candidate decision flows, and we

aggregate process rewards across these candidates using score-weighted Monte Carlo aggregation. To address the path–answer gap, we gate process rewards on answer correctness, so the policy receives process credit only when its final answer is correct. The controller and policy are optimized jointly, and the policy is never conditioned on an oracle decision path. Instead, the policy generates its own decision flow and receives guideline supervision only through scalar process rewards. We evaluate GUIDER on MedGUIDE (Li et al., 2025), an oncology benchmark derived from clinical guidelines with annotated decision paths, and assess transfer to the out-of-domain benchmarks CancerGUIDE (Unell et al., 2025), AMEGA-LLM (Fast et al., 2024), and RealMedQA (Kell et al., 2024).

This paper makes three contributions:

- We formulate guideline decision flows as process supervision for clinical RLVR, accounting for the *path–answer gap* and *path multiplicity* that arise when multiple guideline-consistent decision flows can support the same clinical outcome.
- We introduce a joint controller–policy optimization framework that learns decision flows from guideline-derived supervision and uses multiple controller-generated flow candidates to construct an answer-gated process reward through score-weighted Monte Carlo aggregation.
- We evaluate across MedGUIDE, AMEGA-LLM, RealMedQA, and CancerGUIDE. Our approach achieves 82.5 weighted accuracy and 73.5 guideline adherence on MedGUIDE, outperforming the strongest RLVR baseline by 3.3 and 3.1 points, respectively, while improving AMEGA-LLM to 22.3 and consistently improving RealMedQA and CancerGUIDE.

2 Related Work

Clinical Guideline Benchmarks. Clinical practice guidelines (CPGs) encode evidence-based recommendations for diagnosis and treatment as conditional decision structures, which makes them a natural basis for evaluating clinical reasoning. Recent benchmarks evaluate different aspects of reasoning grounded in guidelines. MedGUIDE derives decision-making tasks from oncology decision trees (Li et al., 2025), and CancerGUIDE evaluates longitudinal treatment trajectories for which several clinical pathways are plausible (Unell et al., 2025). Other benchmarks assess adherence to guidelines in multistage decision processes and open-ended clinical scenarios (Chen et al., 2026; Fast et al., 2024). RealMedQA (Kell et al., 2024) and GuidelineQA (Ding et al., 2026) evaluate realistic clinical questions whose answers depend on guideline recommendations, and Meditron3 (Theimer-Lienhard et al., 2026) provides a large collection of multiple-choice questions derived from the guidelines of multiple healthcare organizations. These benchmarks measure whether a model follows clinical guidelines but do not address how to train models to do so, and benchmarks that annotate decision paths use them only as evaluation targets. We use these decision paths as supervision for reinforcement learning. Because several guideline pathways can be valid for a single patient case, as CancerGUIDE makes explicit, GUIDER treats decision flows as latent variables instead of fixed targets. We evaluate GUIDER across these benchmarks.

Guideline Adherence in LLMs. A complementary line of research represents CPGs in forms that LLMs can retrieve or execute. CPGPrompt transforms guidelines into decision trees that an LLM traverses at inference (Deng et al., 2026), Guideline2Graph extracts clinical decision graphs from guideline documents (Kilic et al., 2026), and Case2Flow retrieves the guideline flowcharts relevant to a patient case (Wei et al., 2026). GuideSkill compiles guideline criteria into executable skills that LLM agents then use (Cao et al., 2026), and MEGA-CDP formulates guideline adherence as a multistage clinical decision process (Chen et al., 2026). Closest to our work, MedGuideX trains LLMs on question answering data generated by executing guideline recommendations compiled into decision trees (Shen et al., 2026). Most of these approaches depend on retrieved, predefined, or executable CPGs at inference time. MedGuideX compiles guideline recommendations into deterministic decision trees, so each case yields a single executed path. The model imitates verbalized traces of this path during supervised fine-tuning, and its reinforcement learning stage rewards only the final answer. Like MedGuideX, GUIDER uses guidelines only during training and requires no guideline access at inference. GUIDER differs in rewarding the reasoning process directly during reinforcement learning. The process reward is aggregated over multiple plausible flows and gated on answer correctness, which accounts for path multiplicity and the path–answer gap.

Verifiable and Process Rewards for Clinical Reasoning. Reinforcement learning with verifiable rewards (RLVR) improves reasoning by optimizing against outcomes that can be checked automatically (Shao et al., 2024; DeepSeek-AI, 2025), and process reward models (PRMs) provide denser supervision over intermediate reasoning steps (Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2024). Recent clinical methods use a range of process rewards. CRPO jointly rewards accuracy, faithfulness, and comprehensiveness (Gu et al., 2025), MRPO introduces step-aware rewards that penalize invalid intermediate reasoning (Jung et al., 2026), and FaithMed evaluates reasoning and retrieval steps against criteria designed by clinicians (Zhang et al., 2026). These process rewards score reasoning traces against fixed criteria and do not use the conditional structure of clinical guidelines. Deriving process rewards from guidelines introduces two challenges. Under path multiplicity, several flows consistent with a guideline can support the same outcome, and a single reference flow would penalize valid alternative decision flows. Under the path–answer gap, a valid flow need not determine the answer for a given patient, so process and outcome rewards can disagree. GUIDER addresses this by generating multiple candidate decision flows from a controller trained with guideline-derived supervision and aggregating process rewards across these candidates. It addresses the path–answer gap by gating this reward on answer correctness, so that only rollouts with a correct final answer receive process credit.

3 METHOD

3.1 PROBLEM FORMULATION

Setup and Notations. Let $x \in \mathcal{X}$ denote a patient case and $y^* \in \mathcal{Y}$ its target clinical answer. We assume that each training pair (x, y^*) is associated with a clinical practice guideline G, which is used to derive a reference decision flow z^* for training supervision but is not provided to the policy at inference. We denote the policy by π_θ, with parameters θ, which generates a decision flow $\hat{z}$ and final answer $\hat{y}$ directly from the patient case, $(\hat{z}, \hat{y}) \sim \pi_\theta(\cdot \mid x)$. We further introduce a decision-flow controller q_ϕ, with parameters ϕ, which generates candidate decision flows and is used only during training to provide process supervision to the policy. Thus, z^* denotes the annotated reference flow, z a controller-generated plausible flow, and $\hat{z}$ the policy-generated flow. At inference, only $\pi_\theta(\cdot \mid x)$ is required; neither G, z^*, nor the controller is needed. Our focus is therefore on learning guideline-aligned decision making rather than guideline retrieval or identification. The controller and policy share model parameters in our implementation, i.e., $\phi = \theta$, and are jointly optimized during training.

Definition 1 (Clinical Decision Flow)**.** *Given a patient case x and a clinical guideline $\mathcal{G}$, a* clinical decision flow *is a sequence:*

$$z = (v_1, v_2, \ldots, v_T), \tag{1}$$

where each v_t represents a guideline-derived clinical condition, criterion, decision, or action. A transition (v_t, v_{t+1}) is valid if it is consistent with both the guideline logic in G and the patient state described by x. We denote the set of valid decision flows for x under guideline G by $\mathcal{Z}(x; G)$.

The learning problem is to train a policy π_θ that, given a patient case x, generates both a decision flow $\hat{z}$ and a final clinical answer $\hat{y}$: $(\hat{z}, \hat{y}) \sim \pi_\theta(\cdot \mid x)$. The final answer should be clinically correct, while the generated decision flow should be consistent with the relevant guideline.

Path multiplicity. Clinical guideline reasoning is generally many-to-one: multiple valid decision flows may support the same clinical answer. For a given patient–outcome pair (x, y^*), there may exist:

$$z_i \neq z_j, \qquad z_i, z_j \in \mathcal{Z}(x; G), \qquad y^* \in \mathcal{Y}(x, z_i) \cap \mathcal{Y}(x, z_j), \tag{2}$$

where $\mathcal{Y}(x, z)$ denotes the set of clinical answers compatible with patient x and decision flow z. Thus, supervision against a single observed or annotated trajectory may incorrectly penalize alternative but clinically valid decision flows.

Definition 2 (Latent Decision Flow)**.** *A decision flow z is* latent *when it is not directly observed from the patient–outcome pair (x, y^*) and is not uniquely determined by it. In particular, the set $\mathcal{Z}(x; G)$ may contain multiple guideline-consistent flows that support the same observed clinical outcome y^*.*

Path–Answer Gap. A guideline-consistent decision flow does not necessarily encode the final answer required by the task. In particular, its terminal node v_T may represent a clinical state, diagnostic test, eligibility criterion, or management action, and therefore need not correspond to the

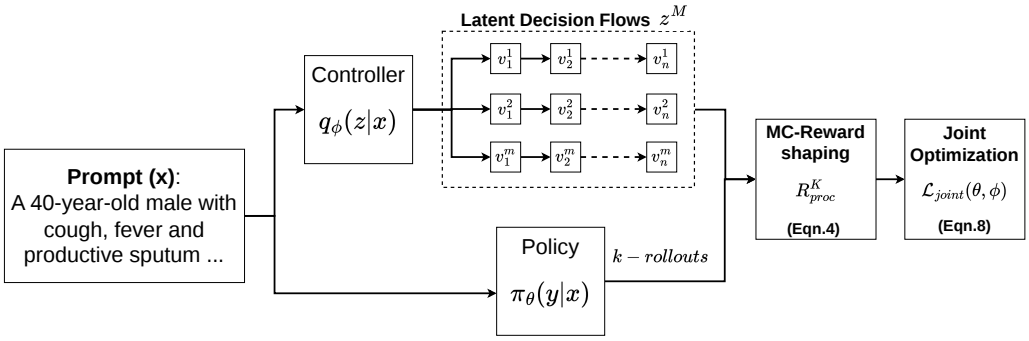


Figure 2: **Overview of GUIDER.** The controller is supervised using guideline-derived decision flows, while the policy independently generates its own decision flow and final answer. Guideline supervision is transferred to the policy through our answer-gated process reward and GRPO, without conditioning the policy on an oracle trajectory.

target answer, $v_T \not\equiv y^*$. Consequently, answer correctness ($\hat{y} = y^*$) and decision-flow validity ($\hat{z} \in \mathcal{Z}(x; G)$) are distinct objectives. A policy may produce the correct answer without following a valid guideline-consistent flow, or generate a valid flow without producing the correct task answer.

Clinical intuition. Consider the chest-pain example in Fig. 1, where the clinical question asks when high-sensitivity troponin testing should be used. A *clinical decision flow* represents the sequence of guideline-derived clinical considerations leading to this decision, such as assessing clinical suspicion of acute coronary syndrome, performing risk stratification, and determining whether troponin testing is indicated. Importantly, different guidelines, such as NICE and SBC, may express distinct but clinically consistent sequences that support the same recommendation, illustrating *path multiplicity*. Given only the patient case and final answer, the particular valid flow underlying that answer is therefore not uniquely determined, making the decision flow *latent*. Finally, neither guideline flow directly states the complete task answer; rather, the answer must be derived from the clinical decisions encoded along the flow, illustrating the *path–answer gap*. Together, these properties motivate accounting for multiple plausible decision flows while jointly optimizing decision-flow validity and final-answer correctness.

3.2 GUIDER APPROACH

As illustrated in Fig. 2, GUIDER operates through a joint controller–policy training framework. Given a patient case x, the controller generates multiple candidate decision flows, while the policy independently generates its own decision flow and final answer. The controller-generated flows are aggregated to construct a process reward for the policy, allowing supervision to account for multiple plausible paths rather than a single reference trajectory. This process reward is combined with answer correctness to optimize the policy, while the controller remains grounded through guideline-derived flow supervision. Through this joint optimization, GUIDER transfers structured guideline supervision to the policy without requiring guideline access at inference time.

We describe GUIDER in three stages: (1) *Decision Flow Controller*, which learns to generate candidate decision flows from guideline-derived supervision; (2) *Multi-Flow Process Reward*, which aggregates process credit across these candidates; and (3) *Joint Controller–Policy Optimization*, which jointly optimizes guideline-consistent reasoning and answer correctness.

3.2.1 DECISION FLOW CONTROLLER

The first component of GUIDER is the decision-flow controller, which learns to generate candidate decision flows from guideline-derived supervision. During training, each case is paired with an annotated reference flow $z^* = (v_1^*, \ldots, v_T^*)$ derived from its associated guideline G. We train the controller autoregressively by minimizing: $\mathcal{L}_{\text{ctrl}}(\theta) = -\sum_{t=1}^{T} \log q_\theta\left(v_t^* \mid x, v_{<t}^*\right)$, where $v_{<t}^* = (v_1^*, \ldots, v_{t-1}^*)$ denotes the reference flow prefix. Importantly, the reference flow z^* supervises the

controller but is not treated as the unique reasoning target for the policy. Instead, the controller learns to generate decision flows from which multiple candidate trajectories can be sampled, allowing subsequent process supervision to account for path multiplicity.

3.2.2 LATENT DECISION-FLOW REWARD

Multiple-Flow Process Supervision. Path multiplicity makes supervision against a single annotated flow overly restrictive. Rather than directly matching the policy-generated flow to z^*, we use the controller to generate multiple candidate decision flows. At training step t, we sample: $z^{(m)} \sim q_{\theta_t}(z \mid x),\ m = 1, \ldots, M$, where θ_t denotes the shared model parameters before the current update.

We first measure the support for each controller proposal using its coverage of the annotated reference flow. Let $\mathcal{V}(z^*)$ denote the reference decision nodes and let $\mathrm{match}(v, z)$ indicate whether reference node v is matched by flow z. We define:

$$s_m = \lambda_{\mathrm{PRM}} \frac{1}{|\mathcal{V}(z^*)|} \sum_{v \in \mathcal{V}(z^*)} \mathrm{match}(v, z^{(m)}), \tag{3}$$

where λ_{PRM} scales the path-coverage score.

We convert these scores into normalized proposal weights: $w_m = \frac{\exp(s_m/\tau)}{\sum_{j=1}^{M} \exp(s_j/\tau)}, \quad m = 1, \ldots, M$ where $\tau > 0$ controls how strongly the reward concentrates on higher-scoring controller proposals.

Monte Carlo Process Reward. To measure agreement between the policy flow and a controller proposal, we use token-set Jaccard similarity. Let $\mathrm{Tok}(z)$ denote the set of alphanumeric tokens obtained after lowercasing, removing common stopwords, and discarding tokens of length at most two. We define: $J(z_1, z_2) = \frac{|\mathrm{Tok}(z_1) \cap \mathrm{Tok}(z_2)|}{|\mathrm{Tok}(z_1) \cup \mathrm{Tok}(z_2)|}$, with $J(z_1, z_2) = 0$ when either token set is empty.

For a policy-generated flow $\hat{z}$, each controller proposal contributes process credit according to both its reference-path support and its agreement with $\hat{z}$. The resulting process reward is:

$$R_{\mathrm{proc}}(\hat{z}) = \sum_{m=1}^{M} w_m \, [s_m]_+ \, J\left(\hat{z}, z^{(m)}\right), \qquad [a]_+ = \max(0, a). \tag{4}$$

This reward can equivalently be viewed as an expectation of proposal-specific process credit under the score-induced categorical distribution; we provide the full derivation in Appendix A.1.

Thus, controller proposals that receive stronger support from the annotated reference flow contribute more heavily, while the policy receives process credit according to its agreement with the resulting weighted set of candidate flows. Unlike direct supervision against z^*, this construction aggregates process credit across multiple controller-generated alternatives.

Answer-Gated Reward Shaping. While multi-flow process supervision accounts for *path multiplicity*, the resulting process reward alone does not address the *path–answer gap*: a guideline-consistent decision flow need not produce the correct clinical answer. We therefore couple the multi-flow process reward with the verifiable answer reward $R_{\mathrm{ans}}(\hat{y}, y^*) = \mathbb{1}[\hat{y} = y^*]$. The reward assigned to a policy-generated completion $(\hat{z}, \hat{y})$ is: $R(\hat{z}, \hat{y}) = R_{\mathrm{ans}}(\hat{y}, y^*) + \lambda R_{\mathrm{ans}}(\hat{y}, y^*) R_{\mathrm{proc}}(\hat{z})$, where $\lambda \geq 0$ controls the contribution of process supervision. The answer gate assigns process credit only when the generated flow also produces the correct clinical answer. Together, multi-flow process supervision and answer-gated reward shaping address the two challenges identified in Sec. 3.1: the former accounts for multiple plausible decision flows, while the latter couples guideline-consistent reasoning with outcome correctness.

3.2.3 JOINT CONTROLLER–POLICY OPTIMIZATION

We integrate controller supervision and policy optimization in a single shared-parameter objective. The controller role is grounded by guideline-derived reference flows, while the policy role is optimized using answer-gated process rewards aggregated over multiple controller-generated flow candidates. Both roles share the same model parameters and are updated jointly.

At training step t, the shared model has parameters θ_t. Using the same pre-update model, we sample M controller flow candidates and independently generate K policy rollouts:

$$z^{(m)} \sim q_{\theta_t}(z \mid x), \qquad (\hat{z}^{(k)}, \hat{y}^{(k)}) \sim \pi_{\theta_t}(\cdot \mid x), \tag{5}$$

for $m = 1, \ldots, M$ and $k = 1, \ldots, K$. Both sets of samples are generated before any parameter update at step t.

For each controller proposal $z^{(m)}$, we compute its reference-path score s_m as defined in Eqs. 3. For each policy rollout k, the process reward is then

$$R^{(k)}_{\text{proc}} = \sum_{m=1}^{M} w_m [s_m]_+ J\left(\hat{z}^{(k)}, z^{(m)}\right), \tag{6}$$

where $J(\cdot, \cdot)$ measures agreement between the policy flow and each controller proposal. The corresponding answer-gated reward is

$$R^{(k)} = R^{(k)}_{\text{ans}} + \lambda R^{(k)}_{\text{ans}} R^{(k)}_{\text{proc}}, \qquad R^{(k)}_{\text{ans}} = \mathbb{1}[\hat{y}^{(k)} = y^*]. \tag{7}$$

Following GRPO, rewards are normalized within the group to obtain $A^{(k)} = \frac{R^{(k)} - \mu_R}{\sigma_R + \epsilon}, \quad k = 1, \ldots, K$ where μ_R and σ_R are the mean and standard deviation of the K rollout rewards. These advantages are used in the standard GRPO objective $\mathcal{L}_{\text{GRPO}}$.

Because the controller and policy share parameters, their objectives are combined before a single optimizer update:

$$\mathcal{L}_{\text{joint}}(\theta_t) = \mathcal{L}_{\text{GRPO}}\left(\theta_t; \{R^{(k)}\}_{k=1}^{K}\right) + \alpha_t \mathcal{L}_{\text{ctrl}}(\theta_t), \tag{8}$$

where α_t controls the contribution of controller supervision. A single optimizer step produces θ_{t+1}. Controller proposals, policy rollouts, and their rewards remain fixed throughout this update, ensuring that both objectives are computed from the same pre-update model θ_t.

Thus, guideline-derived flow supervision and policy optimization jointly shape a single model, while multi-flow process rewards avoid treating the annotated reference flow as the unique reasoning target. At inference time, only the policy $\pi_\theta(\cdot \mid x)$ is used; neither the guideline G, the reference flow z^*, nor controller sampling is required.

4 Experiments

4.1 Experimental Setup

Datasets and Performance Metrics. We train on the MedGUIDE training split (Li et al., 2025) and a subset of Meditron3 (GuidelineQA) (Theimer-Lienhard et al., 2026). We evaluate on held-out MedGUIDE and cross-benchmark on AMEGA-LLM (Fast et al., 2024), RealMedQA (Kell et al., 2024), and CancerGUIDE (Unell et al., 2025).

We follow the evaluation protocol and judge templates provided by each benchmark, using GPT-5.4 as the LLM judge where required. For MedGUIDE, we report weighted accuracy (WAcc.) following Li et al. (2025) and guideline adherence on a 0–5 scale. For AMEGA-LLM, we report the criterion-based aggregate score (0–50). For RealMedQA, we report the mean flow-judge score (0–5), and for CancerGUIDE, the mean clinical-agreement score (0–5). The benchmark-provided evaluation templates and scoring rubrics are reproduced in Appendix A.4.

Baselines. We compare against medical foundation models, recent clinical reinforcement learning methods, and frontier LLMs. The medical models are Meditron-3 (Theimer-Lienhard et al., 2026) and Fastino-Healthcare (Atreja et al., 2026). Our primary baselines are Clinical-R1 (CRPO) (Gu et al., 2025), which uses multi-objective verifiable rewards, MRPO (Jung et al., 2026), which uses step-wise process rewards, and FaithMed (Zhang et al., 2026), which uses rubric-based process supervision for reasoning grounded in evidence. We also report GPT-5.6-Sol and Gemini-Flash-3.7 as frontier references.

Implementation Details. We use Qwen3-14B as our primary backbone and Gemma4-12B as a second backbone. Models are fine-tuned with QLoRA using 4-bit quantization, bfloat16 computation,

Table 1: **Analysis of training and optimization strategies.** We compare the base model, answer-only SFT and GRPO, decision-flow supervision, and GUIDER.

Training Strategy	MedGUIDE		AMEGA-LLM	RealMedQA	CancerGUIDE
	WAcc.	Adh.			
Base (Qwen3-14B)	52.3	47.2	17.2	1.62	1.71
SFT (ans)	78.0	53.1	20.3	1.73	1.73
SFT (ans) → GRPO (ans)	80.1	54.5	20.5	1.81	1.75
SFT (ans+flow)	64.0	71.3	18.8	1.82	1.80
SFT (ans+flow) → GRPO (ans+flow)	68.2	71.6	19.1	1.85	1.82
GUIDER	**82.5**	**73.5**	**22.3**	**1.90**	**2.40**

LoRA rank 64, scaling factor 128, and dropout 0.05. Answer-only SFT is trained for one epoch with learning rate 2×10^{-4}; GRPO and joint optimization are initialized from this checkpoint. For GRPO and joint optimization, we sample $K = 4$ completions, use $\beta = 0.04$, learning rate 5×10^{-6}, batch size 16, maximum prompt length 2048, and disable thinking. Joint-MC policy completions are limited to 256 tokens. For joint optimization, we set $\alpha_t = 1.0$ and $\lambda = 0.2$, and sample $M = 4$ controller flows with temperature 0.7 and top-$p = 0.9$, limited to four decision nodes and 128 tokens. We set the reweighting temperature to $\tau = 0.2$ and use the score-weighted MC process reward in Eq. 4. Joint-MC is trained for one epoch (approximately 655 updates on 10,466 examples) on a single NVIDIA H200 GPU. Evaluation uses deterministic decoding with thinking disabled. Unless otherwise stated, we report the mean and standard deviation over three independent runs with different random seeds. All methods use the same training data and initialization when applicable.

4.2 RESULTS

Table 1 answers three questions about incorporating guideline decision flows into training.

Q1: Can process rewards alone shape RL toward guideline adherence? Only to a limited extent. As shown in Table 1, process-reward optimization yields only modest improvements in guideline adherence (SFT (ans+flow) → GRPO (ans+flow): 71.3→71.6), while answer performance remains substantially below answer-oriented training (68.2 vs. 80.1 WAcc.). This limitation arises because process rewards can only assign useful credit to decision flows that the policy is able to explore; without sufficient in-domain guideline knowledge, the base policy has limited access to high-quality guideline-consistent trajectories. The problem is further compounded by *path multiplicity*: multiple decision flows may be valid for the same clinical outcome, making reward assignment against a single reference flow overly restrictive. Thus, process rewards alone are insufficient for reliably acquiring guideline reasoning from scratch, motivating the use of supervised guideline knowledge to initialize the policy.

Q2: Can we first learn guideline decision flows with SFT and then optimize the resulting policy with RL? A natural solution to the exploration limitation above is to first inject guideline knowledge through supervised fine-tuning. Indeed, flow+answer SFT substantially improves guideline adherence over answer-only SFT (53.1→71.3), demonstrating that the policy can acquire guideline-consistent decision patterns from supervised trajectories. However, this gain comes at a substantial cost in weighted answer accuracy (78.0→64.0). Subsequent GRPO preserves the adherence gain (71.6) but only partially recovers answer accuracy (68.2). This divergence exposes the *path–answer gap*: learning to follow a guideline-consistent decision flow does not necessarily improve the final clinical answer. Moreover, committing the policy to a single supervised trajectory remains problematic under path multiplicity. Thus, sequentially learning flows through SFT and subsequently optimizing with RL does not fully reconcile process quality with outcome correctness.

Q3: Can we jointly optimize guideline reasoning and answer correctness? The limitations above motivate learning the decision process and clinical outcome jointly rather than in separate stages. Our joint MC formulation couples policy optimization with a controller that generates multiple candidate decision flows, allowing process supervision to evolve alongside the policy while accounting for path multiplicity through score-weighted aggregation. This avoids requiring the initial policy to discover high-quality guideline flows on its own and avoids treating a single annotated trajectory as the unique reasoning target. Joint optimization consequently achieves the strongest combina-

Table 2: **Performance across clinical reasoning benchmarks**. For fine-tuned methods, we report mean ± standard deviation over three random seeds. We report weighted accuracy (WAcc.) and guideline adherence (Adh.) on MedGUIDE, aggregate score on AMEGA-LLM, and LLM-judge scores on RealMedQA and CancerGUIDE. Higher is better for all metrics.

Model	MedGUIDE		AMEGA-LLM	RealMedQA	CancerGUIDE
	WAcc.	Adh.			
GPT-5.6-Sol	70.0	67.2	19.1	**2.8**	2.6
Gemini-Flash-3.7	67.2	61.5	18.8	2.4	2.3
Meditron-3-8B (Theimer-Lienhard et al., 2026)	34.7	62.3	18.3	1.5	1.9
Fastino-Healthcare-30B (Atreja et al., 2026)	36.7	41.3	16.1	1.6	1.3
Gemma4-12B					
CRPO (Gu et al., 2025)	62.1±0.6	63.5±0.7	18.2±0.3	1.2±0.1	1.9±0.1
FaithMed (Zhang et al., 2026)	63.5±0.5	55.3±0.8	19.4±0.2	1.3±0.1	2.1±0.1
MRPO (Jung et al., 2026)	63.9±0.6	55.6±0.7	19.3±0.3	1.4±0.1	2.4±0.1
GUIDER	**67.5±0.5**	**65.2±0.6**	**20.2±0.2**	**1.7±0.1**	**2.8±0.1**
Qwen3-14B					
CRPO (Gu et al., 2025)	75.3±0.5	70.4±0.7	21.2±0.2	1.6±0.1	1.9±0.1
FaithMed (Zhang et al., 2026)	77.8±0.4	66.9±0.6	20.7±0.3	1.7±0.1	2.1±0.1
MRPO (Jung et al., 2026)	79.2±0.5	67.5±0.6	21.9±0.2	1.8±0.1	2.3±0.1
GUIDER	**82.5±0.4**	**73.5±0.5**	**22.3±0.2**	**1.9±0.1**	**2.4±0.1**

tion of MedGUIDE weighted accuracy (82.5) and guideline adherence (73.5), while also providing the strongest overall transfer across the evaluated out-of-domain benchmarks. These results support jointly optimizing process and outcome signals while explicitly accounting for the one-to-many relationship between clinical answers and valid guideline decision flows.

Comparison with baselines. Table 2 compares GUIDER against frontier and open-source medical LLMs, as well as recent clinical reasoning optimization methods. GUIDER consistently outperforms CRPO, FaithMed, and MRPO across both model families. On Qwen3-14B, GUIDER achieves the highest MedGUIDE weighted accuracy and guideline adherence (82.5/73.5), improving over the strongest baseline by 3.3 and 3.1 points, respectively. These gains extend beyond the in-domain benchmark: GUIDER also obtains the highest AMEGA-LLM score (22.3) and improves over all three baselines on RealMedQA and CancerGUIDE. A similar pattern holds for Gemma4-12B, where GUIDER improves both answer accuracy and guideline adherence while providing stronger overall out-of-domain performance. Notably, the gains are not explained by using a medically specialized backbone: GUIDER substantially outperforms Meditron-3-8B and Fastino-Healthcare-30B despite operating on general-purpose base models. Finally, on Qwen3-14B, GUIDER surpasses GPT-5.6-Sol on MedGUIDE weighted accuracy (82.5 vs. 70.0) and adherence (73.5 vs. 67.2), while GPT-5.6-Sol remains stronger on RealMedQA (2.8 vs. 1.9). Overall, these results show that the improvements are consistent across model scales and extend from guideline adherence to answer accuracy and out-of-domain clinical reasoning.

5 DISCUSSION

Limitations and Future Work. A primary limitation of GUIDER is that the controller and policy share the same model, motivated by the policy's limited in-domain guideline knowledge. Future work could explore separate controllers with stronger guideline-specific knowledge. Second, sampling multiple controller flows introduces additional computational overhead, motivating more efficient proposal generation and selection. Finally, our experiments focus primarily on oncology guidelines; extension to other specialties and clinical settings remains future work.

Conclusion. We introduce GUIDER, a joint controller–policy optimization framework for guideline-adherent clinical reasoning. GUIDER uses clinical decision flows as process supervision. It accounts for path multiplicity by aggregating process rewards over multiple controller-generated decision flows, and for the path–answer gap by gating this reward on answer correctness. Across in-domain and out-of-domain benchmarks and two model families, GUIDER improves answer accuracy and guideline adherence over answer-only training, supervised flow training, and clinical process reward baselines. These results show that structured clinical guidelines can serve as a training signal for medical reasoning without requiring access to guidelines at inference.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, improving the clarity and presentation of the manuscript, and supporting code development and debugging. Generative AI was not used to autonomously conduct experiments, determine research conclusions, or evaluate the proposed method. All AI-assisted text and code were reviewed by the authors, and experimental results, analyses, citations, and scientific claims were independently verified. The authors take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

GUIDER is a research method and has not been validated for clinical use. All data come from publicly available benchmarks used under their licenses, and this work involves no human subjects and no identifiable patient information. The main risk is that a model trained to follow guidelines may produce fluent but incorrect or outdated recommendations, and our metrics, several of which rely on LLM judges, do not establish clinical safety. The underlying guidelines come mainly from high-income countries and may encode contested demographic factors, and we do not assess fairness across patient subgroups. Clinical deployment would require prospective validation and oversight by clinicians.

REPRODUCIBILITY STATEMENT

We provide an anonymous implementation of GUIDER, including training and evaluation code, configuration scripts, dependencies, and evaluation templates, at `https://anonymous.4open.science/r/GUIDER-4686`. The implementation details and hyperparameters required to reproduce our experiments are described in the main text and appendix. We additionally document the datasets, evaluation protocols, and LLM-based evaluation templates used for each benchmark in the appendix and accompanying code repository.

REFERENCES

Halimat Afolabi, Zainab Afolabi, Elizabeth Friel, Jude Roberts, Antonio Ji-Xu, Lloyd Chen, Egheosa Ogbomo, Emiliomo Imevbore, Phil Eneje, Wissal El Ouahidi, et al. Faithful or just plausible? evaluating the faithfulness of closed-source llms in medical reasoning. *arXiv preprint arXiv:2603.13988*, 2026.

Dhruv Atreja, Julia White, Nikhil Nayak, Kelton Zhang, Henrijs Princis, George Hurn-Maloney, Ash Lewis, and Urchade Zaratiana. Pioneer agent: Continual improvement of small language models in production, 2026. URL `https://arxiv.org/abs/2604.09791`.

Lang Cao, Yuhao Shen, Tianyang Luo, Simo Du, Hao Peng, and Yue Guo. Guideskill: Evolving executable llm agent skills for guideline-grounded clinical reasoning. *arXiv preprint arXiv:2607.26160*, 2026. URL `https://www.alphaxiv.org/abs/2607.26160`.

Nuo Chen, Xinyang Jiang, Zilong Wang, Zhifei Zhang, Xiaoye Qu, Jiajun Deng, Yulan Guo, and Cairong Zhao. Benchmarking clinical decision pathway adherence in large language models. *arXiv preprint arXiv:2608.26592*, 2026. URL `https://www.alphaxiv.org/abs/2608.26592`.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL `https://arxiv.org/abs/2501.12948`.

Ruiqi Deng, Geoffrey Martin, Tony Wang, Gongbo Zhang, Yi Liu, Chunhua Weng, Yanshan Wang, Justin F. Rousseau, and Yifan Peng. Cpgprompt: Translating clinical guidelines into llm-executable decision support. *arXiv preprint arXiv:2601.03475*, 2026. URL `https://arxiv.org/abs/2601.03475`.

Qing Ding, Eric Hua Qing Zhang, Felix Jozsa, and Julia Ive. Building a silver-standard dataset from nice guidelines for clinical llms. In *Opening the Personal Gate between Technology and Health*

Care: Proceedings of MIE 2026, pp. 919–923. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2026.

Saba A. Farahani, Hung Cao, Ramesh Jain, and Amir M. Rahmani. T2d-bench: Evidence-gated evaluation of llm outputs for type 2 diabetes using a multi-layer clinical–lifestyle knowledge graph. *arXiv preprint arXiv:2606.24145*, 2026. URL https://www.alphaxiv.org/abs/2606.24145.

Dennis Fast, Lisa C. Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, Alexander Loser, and Keno K. Bressem. Autonomous medical evaluation for guideline adherence of large language models. *npj Digital Medicine*, 7:358, 2024. doi: 10.1038/s41746-024-01356-6.

Gilson Soares Feitosa-Filho, Luciano Moreira Baracioli, Carlos José Dornas G Barbosa, André Franci, Ari Timerman, Leopoldo Soares Piegas, José Antônio Marin-Neto, and José Carlos Nicolau. Sbc guidelines on unstable angina and non-st-elevation myocardial infarction: executive summary. *Arquivos brasileiros de cardiologia*, 105(3):214–227, 2015.

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. *arXiv preprint arXiv:2512.00601*, 2025. URL https://arxiv.org/abs/2512.00601.

Zhaoyang Jiang, Xuanqi Peng, Fei Teng, Zhizhong Fu, Yunsoo Kim, Jiacong Mi, Zicheng Li, and Honghan Wu. Better accuracies, worse reasoning: A step-level audit of medical chain-of-thought distillation. *arXiv preprint arXiv:2605.28301*, 2026. URL https://www.alphaxiv.org/abs/2605.28301.

Junha Jung, Minbyul Jeong, Suhyeon Lim, Sungwook Jung, Jaehoon Yun, Taeyun Roh, Mujeen Sung, and Jaewoo Kang. Breaking failure cascades: Step-aware reinforcement learning for medical multimodal reasoning. *arXiv preprint arXiv:2606.31825*, 2026. URL https://arxiv.org/abs/2606.31825.

Gregory Kell, Angus Roberts, Serge Umansky, Yuti Khare, Najma Ahmed, Nikhil Patel, Chloe Simela, Jack Coumbe, Julian Rozario, Ryan-Rhys Griffiths, and Iain J. Marshall. Realmedqa: A pilot biomedical question answering dataset containing realistic clinical questions. *arXiv preprint arXiv:2408.08624*, 2024.

Onur Selim Kilic, Yeti Z. Gurbuz, Cem O. Yaldiz, Afra Nawar, Etrit Haxholli, Ogul Can, and Eli Waxman. Guideline2graph: Profile-aware multimodal parsing for executable clinical decision graphs. In *CVPR 2026 MedReasoner Workshop*, 2026. URL https://arxiv.org/abs/2604.02477.

Xiaomin Li, Mingye Gao, Yuexing Hao, Taoran Li, Guangya Wan, Zihan Wang, Yijun Wang, and Xupeng Chen. Medguide: Benchmarking clinical decision-making in large language models. *arXiv preprint arXiv:2505.11613*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri 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.

National Institute for Health and Care Excellence. Recent-onset chest pain of suspected cardiac origin: assessment and diagnosis. NICE Guideline CG95, 2016. Recommendations 1.2.5.1–1.2.5.7, updated 2016.

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

Yuhao Shen, Lang Cao, Simo Du, Yuqing Wang, Juexiao Zhou, Hao Peng, and Yue Guo. Medguidex: Internalizing decision logic from executable guidelines into large language models for clinical reasoning. *arXiv preprint arXiv:2605.26567*, 2026. URL https://www.alphaxiv.org/abs/2605.26567.

Xavier Theimer-Lienhard, Mushtaha El-Amin, Fay Elhassan, Sahaj Vaidya, Victor Cartier-Negadi, David Sasu, Lars Klein, and Mary-Anne Hartley. Fully open meditron: An auditable pipeline for clinical llms, 2026. URL https://arxiv.org/abs/2605.16215.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

Alyssa Unell, Noel C. F. Codella, Sam Preston, Peniel Argaw, Wen-wai Yim, Zelalem Gero, Cliff Wong, Rajesh Jena, Eric Horvitz, Amanda K. Hall, Ruican Rachel Zhong, Jiachen Li, Shrey Jain, Mu Wei, Matthew Lungren, and Hoifung Poon. Cancerguide: Cancer guideline understanding via internal disagreement estimation. In *Proceedings of Machine Learning Research*, volume 297, 2025.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9426–9439, 2024. URL https://arxiv.org/abs/2312.08935.

Jiale Wei, Yufan Chen, Alexander Jaus, Zdravko Marinov, Julian Friedrich, Simon Reiss, Jens Kleesiek, and Rainer Stiefelhagen. Case2flow: Bridging patient cases and guideline flowcharts through multimodal retrieval. *arXiv preprint arXiv:2608.26414*, 2026. URL https://arxiv.org/abs/2608.26414.

Zhiyun Zhang, Liwen Sun, Xiang Qian, and Chenyan Xiong. Faithmed: Training llms for faithful evidence-based medical reasoning. *arXiv preprint arXiv:2607.01440*, 2026. URL https://arxiv.org/abs/2607.01440.

A Appendix

A.1 Derivation of the Multi-Flow Process Reward

At training step t, we draw M candidate decision flows from the controller role of the shared model,

$$z^{(m)} \sim q_{\theta_t}(z \mid x), \qquad m = 1, \ldots, M. \tag{9}$$

Each proposal receives a reference-path coverage score s_m as defined in equation 3. We use these scores to define a categorical distribution over the sampled proposals,

$$w_m = \frac{\exp(s_m/\tau)}{\sum_{j=1}^{M} \exp(s_j/\tau)}, \qquad \sum_{m=1}^{M} w_m = 1. \tag{10}$$

Hence, proposals with greater coverage of the annotated reference flow receive greater mass, while lower-scoring proposals can still contribute to the reward.

For a policy-generated flow $\hat{z}$, define the proposal-specific process credit

$$r_m(\hat{z}) = [s_m]_+ \, J\left(\hat{z}, z^{(m)}\right), \tag{11}$$

where J is the token-set Jaccard similarity defined in equation 4. The process reward is the expectation of this credit under the score-induced categorical distribution over the sampled controller proposals:

$$R_{\text{proc}}(\hat{z}) = \mathbb{E}_{m \sim \text{Cat}(w_1, \ldots, w_M)} \left[r_m(\hat{z}) \right] \tag{12}$$

$$= \sum_{m=1}^{M} w_m [s_m]_+ J\left(\hat{z}, z^{(m)}\right). \tag{13}$$

Table 3: **Effect of the number of sampled decision flows on Qwen3-14B.** We vary the number of controller-sampled flows M used for Monte Carlo process supervision. We report weighted accuracy (WAcc.) and guideline adherence (Adh.) on MedGUIDE, aggregate score on AMEGA-LLM, and LLM-judge scores on RealMedQA and CancerGUIDE. Higher is better for all metrics.

# Decision Flows (M)	MedGUIDE		AMEGA-LLM	RealMedQA	CancerGUIDE
	WAcc.	Adh.			
1	77.1	70.2	20.3	1.6	1.9
2	80.3	72.5	21.6	1.7	2.1
4	82.5	73.5	22.3	1.9	2.4
6	82.4	74.1	22.1	1.9	2.3
8	82.3	74.5	22.2	1.9	2.3

Substituting equation 10 gives the equivalent form

$$R_{\text{proc}}(\hat{z}) = \frac{\sum_{m=1}^{M} \exp(s_m/\tau)\,[s_m]_+\, J\big(\hat{z}, z^{(m)}\big)}{\sum_{j=1}^{M} \exp(s_j/\tau)}. \tag{14}$$

This construction separates two sources of process information: s_m measures the reference-path support of each controller proposal, while $J(\hat{z}, z^{(m)})$ measures its agreement with the policy-generated flow. Together, they provide score-weighted process supervision over multiple controller-generated flow candidates rather than requiring exact agreement with a single annotated reference flow.

A.2 Effect of the Number of Decision Flows

We study the effect of the number of controller-sampled decision flows M used to estimate the process reward. Increasing M from 2 to 4 improves MedGUIDE weighted accuracy from 80.3 to 82.5 and adherence from 72.5 to 73.5, with similar gains across the out-of-domain benchmarks. Beyond $M = 4$, performance largely saturates: weighted accuracy remains stable around 82.3–82.4, while adherence continues to improve modestly to 74.5 at $M = 8$. These results suggest that a small set of sampled flows is sufficient to capture path multiplicity, with additional samples providing diminishing returns.

A.3 Effect of Decision-Flow Length

We study the effect of the maximum number of decision nodes N allowed in each controller-sampled flow. Increasing N from 2 to 4 substantially improves both MedGUIDE weighted accuracy (77.2 to 82.3) and guideline adherence (68.5 to 74.5), with corresponding gains on all out-of-domain benchmarks. Increasing the flow length beyond four nodes provides little additional benefit: adherence remains comparable at $N = 6$ (74.4), while weighted accuracy slightly decreases to 81.2. At $N = 8$, both metrics decline further, with similar degradation on the out-of-domain benchmarks. These results suggest that moderately sized decision flows provide sufficient structure to capture clinically relevant decision steps, while overly short flows omit useful process information and longer flows introduce unnecessary complexity without consistent performance gains.

A.4 Evaluation Templates

For open-ended and guideline-adherence evaluation, we follow the evaluation templates and scoring rubrics provided by the respective benchmarks. We do not modify the benchmark-specific evaluation criteria; we instantiate the LLM judge with GPT-5.4. The templates used in our experiments are reproduced below for completeness. Judges receive the benchmark-specified case, reference evidence or criteria, and model output, but not the model identity or training condition. The templates for MedGUIDE guideline-adherence, AMEGA-LLM, RealMedQA, and CancerGUIDE evaluation are provided in Figs. 5, 6, 7, and 8, respectively.

Table 4: **Effect of the number of decision nodes in sampled decision flows on Qwen3-14B.** We vary the maximum number of decision nodes N in each controller-sampled flow and evaluate its effect on in-domain and out-of-domain performance. We report weighted accuracy (WAcc.) and guideline adherence (Adh.) on MedGUIDE, aggregate score on AMEGA-LLM, and LLM-judge scores on RealMedQA and CancerGUIDE. Higher is better for all metrics.

# Decision Nodes (N)	MedGUIDE		AMEGA-LLM	RealMedQA	CancerGUIDE
	WAcc.	Adh.			
2	77.2	68.5	20.3	1.5	1.8
4	82.3	74.5	22.2	1.9	2.3
6	81.2	74.4	22.4	1.8	2.4
8	80.5	72.8	21.5	1.8	2.1

Controller Prompt

SYSTEM:
You are a clinical decision-flow controller. Given a patient vignette, generate a compact decision flow that will answer the question.

USER:
Question: {question}

Generate the likely guideline-style decision flow. Use exactly this format:
Decision flow:
1. <short decision state>
2. <short decision state>

Figure 3: **LLM instructions for the decision-flow controller.** Given a clinical question, the controller generates a compact guideline-style decision flow. The generated flows represent plausible clinical decision pathways and are used to provide process-level supervision during policy optimization.

A.5 Controller and Policy Prompts

We use separate prompts for the controller and policy, provided in Figures 3 and 4. Both roles receive the same clinical question. The controller is prompted to generate plausible guideline-style candidate decision flows, while the policy independently generates its own decision flow followed by the final answer. The policy does not observe the controller-generated flows; instead, they provide process-level supervision through the reward. This separation provides process supervision without directly constraining the policy to reproduce a particular trajectory.

A.6 Qualitative Analysis

In the MedGUIDE example (Figure 9), the model identifies the key decision states—relapse after systemic therapy and poor performance status (ECOG PS 3)—and generates a concise decision flow consistent with the gold guideline recommendation, ultimately selecting the correct treatment option. This demonstrates that the model can recover clinically relevant decision states without requiring exact reproduction of the reference flow.

Policy Prompt

SYSTEM:
You are taking a clinical multiple-choice benchmark. First generate a compact guideline-style decision flow, then give the final option letter.

USER:
Question: {question}
Your response must use exactly this format:

Decision flow:
1. <short decision state>
2. <short decision state>
Answer: <single option letter>

Use at most N decision-flow nodes. Do not include guideline path text from memory; infer the flow from the vignette.

Figure 4: **LLM instructions for the clinical reasoning policy.** Given a clinical question, the policy generates its own compact guideline-style decision flow followed by the final answer. The policy does not observe controller-generated flows, allowing it to independently reason over the clinical case while receiving process-level supervision through the optimization objective. N denotes the maximum number of decision-flow nodes.

Template

Evaluate whether the model-generated clinical decision flow adheres to the gold MedGUIDE guideline path for this case.

Judge semantic clinical equivalence and ordering, not exact wording. Do not reward generic guideline-like statements unless they correspond to a gold path step. Do not grade MCQ answer correctness as the main criterion; focus on whether the decision flow uses the right clinical criteria and reaches a path compatible with the gold trajectory.

Score rubric:
5 = all key gold decision states are captured in a clinically equivalent order, with no important contradictions.
4 = most key states are captured; only minor omissions or mild ambiguity.
3 = partially aligned but misses an important criterion, branch, or action.
2 = weak or mostly generic alignment; little specific path adherence.
1 = mostly wrong path or clinically contradicted by the gold path.
0 = empty or unusable decision flow.

Return **JSON**: "score": ;

Figure 5: **MedGUIDE Guideline Adherence Judge Template.** For each MedGUIDE example, the judge receives the patient vignette, the supervised gold decision path, the gold answer option, the model's parsed answer, and the model-generated decision flow.

Template

Evaluate whether the model answer satisfies each AMEGA scoring criterion.

Rules:
- Judge only the model answer against the criterion and the case/question.

- Return true only when the answer clearly satisfies the criterion.
- Return false for missing, contradicted, or ambiguous information.
- Do not award credit for unsafe or clinically inappropriate statements.
Return exactly one Boolean verdict per criterion: "verdicts": [true, false, ...]
Each true verdict receives the criterion's predefined weight. We report the aggregate score on a 0–50 scale:

$$\text{AMEGA}_{50} = 50 \frac{\sum_i s_i}{\sum_i s_i^{\max}}. \tag{15}$$

Return **JSON**: "score": ;

Figure 6: **AMEGA judge template.** AMEGA-LLM provides a set of weighted clinical criteria for each case–question pair. The judge evaluates each criterion independently.

Template

You evaluate whether a model-generated clinical decision/support flow is supported by a NICE guideline recommendation. Use only the provided clinical question, reference recommendation, and model output.
Score the model output against the reference NICE recommendation.

Scoring rubric:
5 = fully answers the question and is faithful to the recommendation.
4 = mostly faithful, with only minor omissions.
3 = partially supported or incomplete.
2 = weakly related, with important missing or unsupported components.

1 = mostly unsupported or misleading.
0 = contradicted, unsafe, or unrelated.
Also judge whether the decision flow explicitly connects the clinical question to the recommendation.

Return **JSON**: "score": ;

Figure 7: **RealMedQA judge template.** For RealMedQA, the judge compares the generated clinical decision flow with the reference NICE recommendation.

Template

You evaluate NSCLC treatment recommendation predictions against a reference CancerGUIDE recommendation. Use only the patient profile, reference recommendation, and model output. Score whether the model's final recommendation is clinically equivalent to the reference recommendation.

Scoring rubric:
5 = fully answers the question and is faithful to the recommendation.
4 = mostly faithful, with only minor omissions.
3 = partially supported or incomplete.
2 = weakly related, with important missing or unsupported components.

1 = mostly unsupported or misleading.
0 = contradicted, unsafe, or unrelated.
Also judge whether the decision flow explicitly connects the clinical question to the recommendation.

Return **JSON**: "score": ;

Figure 8: **CancerGuide judge template.** For CancerGUIDE, the judge evaluates whether the generated cancer-treatment recommendation is clinically equivalent to the reference recommendation.

Example

A 72-year-old woman with extensive-stage small-cell lung cancer initially responded to cisplatin/etoposide but relapsed after six months. Second-line topotecan also failed after four months. She now has progressive fatigue, weight loss, dyspnea, new liver and bone metastases, worsening liver function, and ECOG performance status 3.

What treatment option should be considered?
A) Palliative and best supportive care, with consideration of single-agent chemotherapy or palliative radiotherapy
B) Continue systemic therapy
C) Clinical trial

D) Continued surveillance
E) Palliative and best supportive care
F) Palliative/supportive care with possible chemotherapy, targeted therapy, or radiotherapy
G) Adjuvant therapy
Correct answer:

A — Palliative and best supportive care, with consideration of single-agent chemotherapy or palliative radiotherapy.

Gold guideline recommendation:

Poor performance status. Palliative and best supportive care ± single-agent chemotherapy/palliative radiotherapy

Generated decision flow:

1. Relapse or primary progressive disease after systemic therapy for ES-SCLC
2. Poor performance status (ECOG PS 3)
Final answer: A

Figure 9: **Example clinical question and generated decision flow from MedGUIDE.** Given a patient with relapsed extensive-stage small-cell lung cancer and poor performance status, the model generates a compact decision flow identifying the clinically relevant guideline conditions and arrives at the correct treatment recommendation. The generated flow is consistent with the gold guideline recommendation while expressing it as a concise sequence of decision states.